PROTECCIÓN CORPORATIVA
CONTRA DEEPFAKES
MANUAL DE PREVENCIÓN Y RESPUESTA ESTRATÉGICA
ALERT
FAKE
Autor: Lic. Mauricio Alan Illera Díaz
ORCID ID: 0009-0001-1505-4170
Researcher ID: OSI-4768-2025
ID del autor: 9514252
ID Científico: 2A1E-218D-9A52 - NoblePress: PUB1108553

Temario

1. Introducción

1.1 Objetivo del manual
1.2 Alcance (corporativo, filiales, proveedores y terceros)
1.3 Definición de deepfake y deepfake en tiempo real
1.4 Principio clave: *la imagen y la voz ya no validan identidad*
1.5 Riesgos estratégicos para la organización

2. Marco Conceptual y Amenaza Actual

2.1 Evolución del fraude digital
2.2 Ingeniería social avanzada con IA
2.3 Diferencias entre deepfake tradicional y en tiempo real
2.4 Casos reales documentados (finanzas, RH, alta dirección)
2.5 Tendencias y proyección de riesgo

3. Impacto Organizacional y Psicológico

3.1 Psicología del engaño y sesgos cognitivos
3.2 Autoridad, urgencia y manipulación emocional
3.3 Impacto en confianza interna y cultura organizacional
3.4 Daño reputacional y estrés post-incidente
3.5 Gestión emocional en incidentes críticos

4. Marco Legal y Cumplimiento (México)

4.1 Fraude y usurpación de identidad
4.2 Delitos informáticos aplicables
4.3 Protección de datos personales (LFPDPPP)
4.4 Valor probatorio de video y voz
4.5 Responsabilidad corporativa por omisión
4.6 Buenas prácticas de defensa legal preventiva

5. Gobierno Corporativo y Responsabilidades

5.1 Consejo de Administración
5.2 CEO y Alta Dirección
5.3 Seguridad Corporativa
5.4 Auditoría Interna
5.5 Cumplimiento / Legal
5.6 Recursos Humanos
5.7 TI / Ciberseguridad
5.8 Proveedores y terceros

6. Clasificación de Riesgos por Área

6.1 Finanzas y Tesorería
6.2 Compras y Contrataciones
6.3 Recursos Humanos
6.4 Seguridad y Crisis
6.5 Comunicación Corporativa
6.6 Jurídico y Compliance

7. Controles Preventivos (Antes del Incidente)

7.1 Principios de verificación de identidad
7.2 Prohibiciones críticas (qué NO se debe hacer)
7.3 Autorizaciones de alto impacto
7.4 Verificación multicanal
7.5 Frases o códigos dinámicos
7.6 Gestión de accesos y privilegios
7.7 Capacitación y concientización

8. Detección de Deepfake en Tiempo Real

8.1 Indicadores técnicos
8.2 Indicadores conductuales
8.3 Señales de alerta psicológica
8.4 Latencia, sincronía y patrones anómalos
8.5 Reporte inmediato de sospecha

9. Protocolo de Respuesta Inmediata

9.1 Activación del protocolo
9.2 Aislamiento de la decisión
9.3 Verificación por canal seguro
9.4 Escalamiento jerárquico
9.5 Registro y preservación de evidencia
9.6 Comunicación interna controlada

10. Gestión de Crisis y Continuidad del Negocio

10.1 Comité de crisis
10.2 Toma de decisiones bajo incertidumbre
10.3 Manejo reputacional
10.4 Comunicación externa y medios
10.5 Coordinación con autoridades
10.6 Recuperación post-incidente

11. Tecnología y Herramientas de Apoyo

11.1 Sistemas de detección de voz sintética
11.2 Plataformas seguras de comunicación
11.3 Análisis biométrico conductual

11.4 Limitaciones tecnológicas actuales
11.5 Evaluación y selección de proveedores

12. Auditoría, Monitoreo y Mejora Continua

12.1 Indicadores clave (KPIs)
12.2 Auditorías internas
12.3 Pruebas y simulacros
12.4 Actualización del manual
12.5 Lecciones aprendidas

13. Capacitación y Cultura Organizacional

13.1 Programas para directivos
13.2 Entrenamiento operativo
13.3 Simulaciones de fraude
13.4 Cultura de reporte sin represalias
13.5 Evaluación de madurez organizacional

14. Integración con Otros Sistemas

14.1 Seguridad Corporativa
14.2 SHE / Prevención
14.3 Gestión de Riesgos
14.4 Compliance y Ética
14.5 Continuidad del Negocio

15. Anexos

15.1 Glosario
15.2 Checklist anti-deepfake
15.3 Flujogramas de decisión
15.4 Formatos de reporte
15.5 Roles y contactos críticos

1. Introducción

La acelerada adopción de tecnologías basadas en inteligencia artificial ha generado beneficios significativos para la eficiencia, la comunicación y la toma de decisiones en las organizaciones. Sin embargo, este mismo avance ha dado origen a nuevas amenazas de alto impacto estratégico, entre ellas los *deepfakes*, particularmente los deepfakes en tiempo real, los cuales representan un punto de quiebre en la manera en que las empresas validan la identidad, la autoridad y la autenticidad de la información.

Históricamente, la imagen y la voz han sido consideradas pruebas suficientes de identidad. En el contexto actual, esta premisa ha dejado de ser válida. Hoy, una organización puede ser vulnerada no por fallas tecnológicas, sino por confianza mal depositada en estímulos visuales y auditivos que pueden ser completamente falsos.

Este manual surge como una herramienta estratégica de gobierno corporativo, diseñada para proteger a la organización frente a riesgos emergentes que afectan directamente la continuidad del negocio, la integridad financiera, la reputación institucional y la responsabilidad legal de la alta dirección.

1.1 Objetivo del manual

El objetivo de este manual es establecer un marco claro, obligatorio y transversal para la prevención, detección y respuesta ante el uso de deepfakes —especialmente en tiempo real— dentro de la organización.

De manera específica, este manual tiene como objetivos:

- Definir de forma inequívoca qué es un deepfake y por qué constituye un riesgo estratégico.
- Eliminar cualquier ambigüedad sobre la validez de la imagen y la voz como mecanismos de autenticación.
- Establecer principios rectores para la toma de decisiones en escenarios de posible manipulación.
- Proteger a la organización frente a fraudes financieros, suplantación de identidad, espionaje corporativo y daño reputacional.
- Proveer a la alta dirección de una defensa preventiva documentada, alineada a buenas prácticas de cumplimiento, auditoría y responsabilidad fiduciaria.
- Integrar este riesgo emergente dentro del sistema global de gestión de riesgos, seguridad corporativa y continuidad del negocio.

Este manual no es un documento informativo; es un instrumento normativo. Su cumplimiento es obligatorio y su omisión puede derivar en responsabilidades administrativas, operativas y legales.

1.2 Alcance (corporativo, filiales, proveedores y terceros)

El alcance de este manual es total y extensivo, abarcando todas las entidades, personas y relaciones que interactúan con la organización.

Este manual aplica a:

A nivel corporativo

- Consejo de Administración
- CEO y Alta Dirección
- Directores, gerentes y mandos medios
- Colaboradores de todas las áreas

A nivel organizacional ampliado

- Filiales, subsidiarias y unidades de negocio
- Oficinas regionales y representaciones
- Proyectos temporales y equipos especiales

A nivel externo

- Proveedores estratégicos
- Consultores
- Auditores externos
- Socios comerciales
- Terceros con acceso a información, sistemas o procesos críticos

Cualquier interacción que implique:

- Autorización de pagos
- Validación de identidad
- Acceso a información sensible
- Toma de decisiones críticas
- Comunicación ejecutiva

queda automáticamente sujeta a los lineamientos de este manual, independientemente del medio utilizado (videollamada, llamada telefónica, mensaje de voz, transmisión en vivo o plataforma digital).

1.3 Definición de deepfake y deepfake en tiempo real

Deepfake

Un *deepfake* es un contenido audiovisual generado o manipulado mediante inteligencia artificial que simula de manera realista la apariencia, la voz o el comportamiento de una persona, con el objetivo de hacer creer que dicha persona dijo, hizo o autorizó algo que en realidad nunca ocurrió.

Deepfake en tiempo real

Un deepfake en tiempo real es una evolución crítica de esta tecnología. Se trata de sistemas capaces de:

- Modificar rostros, gestos y expresiones en vivo.
- Clonar y reproducir voces con alta fidelidad de forma instantánea.
- Sincronizar labios, tono emocional y lenguaje corporal durante una interacción real.

Esto significa que la manipulación ocurre durante la comunicación, sin necesidad de edición posterior, lo que elimina los indicadores tradicionales de fraude.

En términos ejecutivos:

Un deepfake en tiempo real permite que una persona inexistente tome decisiones reales dentro de la organización.

1.4 Principio clave: la imagen y la voz ya no validan identidad

Este manual se fundamenta en un principio rector que debe ser comprendido y adoptado por toda la organización:

La imagen y la voz, por sí solas, ya no constituyen prueba válida de identidad, autoridad ni intención.

Este principio rompe con décadas de prácticas corporativas basadas en la confianza visual y auditiva. En el entorno actual:

- Ver al CEO no garantiza que sea el CEO.
- Escuchar una voz conocida no garantiza que pertenezca a la persona legítima.
- La urgencia y la presión emocional son indicadores de riesgo, no de legitimidad.

Cualquier proceso, autorización o decisión que dependa exclusivamente de imagen o voz se considera intrínsecamente vulnerable y debe ser rediseñado.

Este principio es no negociable y prevalece sobre jerarquía, antigüedad o contexto.

1.5 Riesgos estratégicos para la organización

Los deepfakes en tiempo real no representan un riesgo tecnológico aislado, sino un riesgo estratégico de nivel Consejo.

Principales riesgos:

Riesgo financiero

- Transferencias no autorizadas
- Fraude en pagos y contratos
- Manipulación de decisiones de tesorería

Riesgo reputacional

- Comunicaciones falsas atribuidas a la empresa
- Declaraciones públicas manipuladas
- Pérdida de confianza de inversionistas y clientes

Riesgo legal y de cumplimiento

- Responsabilidad por omisión de controles
- Invalidación de pruebas audiovisuales
- Sanciones por violación de datos personales

Riesgo operativo

- Accesos indebidos a sistemas críticos
- Sabotaje de procesos clave
- Interrupción de operaciones

Riesgo humano y cultural

- Erosión de la confianza interna
- Clima organizacional de sospecha
- Estrés y errores bajo presión

c

Ignorar este riesgo no es neutral. No prepararse es una decisión estratégica con consecuencias directas sobre la viabilidad y resiliencia de la organización.

Este manual posiciona a la empresa en un enfoque proactivo, responsable y alineado a las mejores prácticas internacionales, asegurando que la alta dirección pueda demostrar diligencia debida frente a una de las amenazas más complejas de la era digital.

2. Marco Conceptual y Amenaza Actual

El riesgo asociado a los deepfakes en tiempo real no puede entenderse como un fenómeno aislado ni como una moda tecnológica. Es el resultado lógico de décadas de evolución del fraude, combinado con el acceso masivo a inteligencia artificial avanzada. Para la alta dirección, comprender este marco conceptual es indispensable, ya que permite dimensionar por qué los controles tradicionales han dejado de ser suficientes y por qué este riesgo debe gestionarse al más alto nivel de gobierno corporativo.

Este capítulo establece una lectura ejecutiva clara: la amenaza ya está activa, es silenciosa, escalable y altamente rentable para el crimen organizado, y se dirige de manera prioritaria a quienes tienen autoridad, poder de decisión y acceso a recursos.

2.1 Evolución del fraude digital

El fraude siempre ha evolucionado en función de tres variables: tecnología disponible, comportamiento humano y controles existentes. Cada vez que las organizaciones fortalecen un control, el fraude migra hacia el punto de menor resistencia.

Fases de evolución del fraude

Primera etapa – Fraude básico
Suplantación simple, documentos falsos, llamadas telefónicas sin sofisticación. Dependía casi exclusivamente del descuido humano.

Segunda etapa – Fraude digital inicial
Correos electrónicos falsos, phishing, malware, robo de contraseñas. El foco estaba en vulnerar sistemas.

Tercera etapa – Fraude dirigido
Spear phishing, fraude al CEO, ataques personalizados. Se comienza a estudiar a la víctima.

Cuarta etapa – Fraude cognitivo y emocional (actual)
Deepfakes, clonación de voz, manipulación en tiempo real. El objetivo ya no es el sistema, sino **la percepción humana de la realidad**.

En esta última etapa, el fraude deja de ser técnico y se convierte en psicológico, estratégico y organizacional. La tecnología solo es el habilitador; el verdadero vector de ataque es la confianza.

Desde una perspectiva de Consejo:

El fraude moderno no rompe firewalls; rompe certezas.

2.2 Ingeniería social avanzada con IA

La ingeniería social tradicional se basaba en engaños genéricos. La ingeniería social avanzada con IA se basa en modelos de comportamiento humano.

¿Qué cambia con la IA?

- Análisis de redes sociales, entrevistas, conferencias y videos públicos.
- Modelado de estilo de comunicación, tono emocional y lenguaje corporal.
- Simulación de jerarquía, urgencia y contexto organizacional.
- Adaptación en tiempo real a las reacciones de la víctima.

La IA permite que el ataque:

- Sea personalizado
- Escale rápidamente
- Se ajuste dinámicamente durante la interacción

En términos prácticos, la víctima no interactúa con un estafador, sino con un sistema entrenado para parecer legítimo.

Factores psicológicos explotados

- Autoridad (CEO, director, consejero)
- Urgencia (“es ahora o perdemos la oportunidad”)
- Confidencialidad (“no involucres a nadie más”)
- Miedo a fallar
- Deseo de cumplir

La ingeniería social con IA no busca errores técnicos; busca obediencia racional bajo presión.

2.3 Diferencias entre deepfake tradicional y deepfake en tiempo real

Comprender esta diferencia es crítico para no subestimar la amenaza.

Deepfake tradicional

- Contenido editado posteriormente
- Uso en videos pregrabados
- Posibilidad de análisis forense
- Menor presión temporal
- Uso frecuente en desinformación

Deepfake en tiempo real

- Manipulación durante la interacción
- Sin edición posterior
- Decisiones tomadas en vivo
- Presión psicológica inmediata
- Difícil detección humana

Desde una óptica ejecutiva:

El deepfake tradicional daña la reputación; el deepfake en tiempo real toma decisiones por la empresa.

Este cambio convierte a las videollamadas, llamadas telefónicas y transmisiones en vivo en superficies de ataque críticas.

2.4 Casos reales documentados (finanzas, RH, alta dirección)

Finanzas y Tesorería

- Transferencias millonarias autorizadas tras videollamadas falsas
- Clonación de voz de directores financieros
- Instrucciones urgentes de pago con narrativa creíble

Recursos Humanos

- Solicitudes falsas de nómina confidencial
- Autorización de despidos o pagos especiales
- Acceso indebido a datos personales

Alta Dirección

- Suplantación de CEO o consejeros
- Manipulación de decisiones estratégicas
- Uso de identidad ejecutiva para influir en terceros

En todos los casos documentados se repite un patrón:

- Contexto creíble
- Autoridad aparente
- Urgencia emocional

- Falta de verificación independiente

Ninguna de las organizaciones afectadas carecía de tecnología; carecían de protocolos para una realidad donde la imagen y la voz ya no son confiables.

2.5 Tendencias y proyección de riesgo

Tendencias claras

- Democratización de herramientas de deepfake
- Reducción de costos de ataque
- Incremento de calidad audiovisual
- Uso por crimen organizado
- Ataques dirigidos a ejecutivos

Proyección estratégica

En los próximos años:

- El fraude por deepfake superará al phishing tradicional
- Las decisiones críticas en remoto serán objetivo prioritario
- El video dejará de ser evidencia suficiente
- Las empresas sin protocolos serán consideradas negligentes

Visión ejecutiva

Este riesgo no se gestiona con una herramienta aislada. Se gestiona con liderazgo, cultura, procesos claros y decisiones firmes.

Prepararse hoy no es una ventaja competitiva; es un requisito de supervivencia corporativa.

Este capítulo establece la base conceptual para justificar controles estrictos, cambios culturales y una nueva forma de ejercer la autoridad en la era digital.

Introducción al impacto organizacional y psicológico

En la era digital, los deepfakes en tiempo real representan un fenómeno disruptivo que trasciende los límites de la tecnología y la seguridad informática, penetrando en el núcleo psicológico y cultural de las organizaciones. Su relevancia no solo radica en la capacidad de falsear identidades y mensajes, sino en el potencial para desestabilizar la toma de

decisiones, la confianza interna y la reputación corporativa. Hoy, la alta dirección enfrenta el reto de comprender y gestionar estos riesgos, reconociendo que la amenaza de los deepfakes es, ante todo, una amenaza humana y organizacional. El contexto actual demanda una reflexión profunda: ¿Cómo proteger la integridad emocional y cultural de la empresa frente a manipulaciones digitales cada vez más sofisticadas?

Psicología del engaño y sesgos cognitivos

El éxito de los deepfakes en tiempo real no depende de deficiencias técnicas, sino de la explotación sistemática de sesgos cognitivos universales. Estos mecanismos psicológicos, presentes en todos los individuos sin distinción de rango o experiencia, son el verdadero campo de batalla para los ataques de ingeniería social avanzada. A continuación, se detallan los principales sesgos cognitivos explotados por los deepfakes, acompañados de ejemplos prácticos y sus implicaciones en la toma de decisiones empresariales.

Sesgo de autoridad: La tendencia automática a obedecer instrucciones de figuras jerárquicas o percibidas como legítimas. Por ejemplo, un gerente recibe una videollamada de un “CEO” solicitando una transferencia urgente; el instinto de obediencia prevalece sobre la verificación, incluso en ambientes de alta seguridad. Este sesgo puede llevar a la ejecución de acciones críticas sin el debido escrutinio.

Sesgo de familiaridad: La voz y el rostro conocidos generan confianza instantánea. Un empleado que reconoce a su supervisor en una videollamada tiende a bajar la guardia, ignorando señales sutiles de manipulación. El deepfake aprovecha esta confianza para facilitar el acceso a información o recursos sensibles.

Sesgo de urgencia: Bajo presión temporal, el cerebro prioriza la acción sobre la verificación, especialmente si la solicitud implica consecuencias inmediatas. Por ejemplo, la narrativa de “tenemos solo cinco minutos antes de perder el contrato” obliga a una reacción instintiva, minimizando el análisis crítico.

Sesgo de coherencia: Si el mensaje encaja con la personalidad y estilo habitual del emisor, se reduce el cuestionamiento. Un deepfake que imita el tono y las expresiones de un líder puede pasar desapercibido, incluso ante equipos entrenados.

Sesgo de responsabilidad difusa: Cuando la instrucción proviene “de arriba”, el individuo siente menor responsabilidad personal y asume que la decisión está validada por la jerarquía. Este sesgo facilita la ejecución de órdenes sin el filtro ético o legal correspondiente.

La inteligencia artificial detrás de los deepfakes no improvisa estos sesgos; los modela y los explota deliberadamente, generando entornos donde incluso los profesionales más preparados pueden ser manipulados. Las implicaciones son profundas: la toma de decisiones bajo presión y en contextos de incertidumbre se vuelve altamente vulnerable, exigiendo la implementación de protocolos que contemplen no solo la seguridad técnica, sino la fortaleza psicológica del personal.

Autoridad, urgencia y manipulación emocional

La manipulación emocional en ataques de deepfake combina tres elementos clave que, juntos, neutralizan los mecanismos racionales de defensa y favorecen la obediencia racionalizada. El uso de identidades ejecutivas reales, acompañado de lenguaje jerárquico convincente y un tono emocional congruente con el liderazgo, crea un ambiente de autoridad indiscutible. A esto se suma la urgencia artificial, con ventanas de tiempo cortas y narrativas de crisis inminente que infunden temor a consecuencias negativas si no se actúa de inmediato.

Casos reales demuestran que empleados de todos los niveles han sido persuadidos para realizar transferencias bancarias, compartir información confidencial o modificar procesos internos bajo la presión emocional de una videollamada deepfake. La amenaza no reside únicamente en el contenido del mensaje, sino en la forma en que se manipulan emociones como el miedo a fallar, el deseo de lealtad y la confidencialidad forzada (“esto queda entre nosotros”).

La conducta organizacional se ve afectada de manera significativa: cuestionar la legitimidad de una orden parece irresponsable, y la urgencia emocional se convierte en un indicador de riesgo más que en una señal de autenticidad. La alta dirección debe reconocer que el verdadero peligro no es la tecnología, sino la capacidad de los atacantes para explotar los instintos humanos en momentos críticos.

Impacto en confianza interna y cultura organizacional

Un solo incidente de deepfake puede desencadenar consecuencias inmediatas y estructurales que alteran las dinámicas de confianza dentro de la organización.

Efectos inmediatos: Tras el incidente, surgen dudas sobre la veracidad de las comunicaciones internas, miedo a tomar decisiones y paralización operativa. Los empleados pueden sentirse inseguros, temiendo que cualquier mensaje podría ser falso.

Efectos estructurales: Se erosiona la confianza vertical entre líderes y equipos, incrementan los controles informales y emerge una cultura de sospecha. La organización transita de una cultura de confianza a una de desconfianza silenciosa, donde las personas evitan asumir responsabilidades, difieren decisiones críticas y operan bajo esquemas de autoprotección, afectando la agilidad, la innovación y el liderazgo efectivo.

Las dinámicas de confianza y desconfianza requieren estrategias de mitigación que incluyan la implementación de protocolos claros, capacitación continua y la promoción de una cultura de reporte abierto. La gestión adecuada del riesgo implica no solo prevenir incidentes, sino restaurar la confianza a través de la transparencia, el apoyo emocional y el liderazgo visible.

Daño reputacional y estrés post-incidente

El daño reputacional derivado de un deepfake trasciende la pérdida financiera directa. Una declaración falsa atribuida a la empresa puede comprometer la credibilidad de la alta dirección y generar desconfianza entre inversionistas, clientes y reguladores. La

exposición mediática y la viralización en redes sociales amplifican el impacto, dificultando la recuperación de la imagen corporativa.

Internamente, las personas involucradas suelen experimentar culpa, ansiedad, pérdida de confianza en su propio juicio y temor a represalias. Si no se gestiona adecuadamente, este estrés puede derivar en rotación de talento clave, bajo desempeño y ocultamiento de errores futuros, agravando el daño y perpetuando una cultura de miedo y silencio.

La prevención de la rotación de talento y el deterioro del clima laboral exige la implementación de programas de apoyo psicológico, espacios de diálogo y la validación pública de la actuación responsable de los empleados involucrados. La gestión del estrés post-incidente es tan importante como la gestión técnica del evento.

Gestión emocional en incidentes críticos

La intervención en casos de deepfake debe centrarse tanto en el control del daño como en la contención emocional y organizacional. Los principios clave incluyen:

No culpabilizar: El error es sistémico, no individual. Evitar la búsqueda de responsables favorece el aprendizaje colectivo y la resiliencia organizacional.

Transparencia controlada: Informar sobre el incidente sin generar pánico, equilibrando la necesidad de conocimiento con la protección emocional de los equipos.

Liderazgo visible: La alta dirección debe dar la cara, mostrando empatía, firmeza y compromiso con la recuperación.

Apoyo psicológico: Facilitar acceso a servicios de orientación y acompañamiento emocional para quienes lo requieran.

El CEO desempeña un rol fundamental, no solo en la respuesta operativa, sino en la narrativa interna. Su mensaje debe reafirmar la confianza institucional, reconocer la complejidad del riesgo y validar la actuación responsable de quienes detectan o reportan sospechas de deepfake. La comunicación interna efectiva es la clave para transformar la adversidad en una oportunidad de aprendizaje y fortalecimiento cultural.

Conclusión ejecutiva

Los deepfakes en tiempo real desafían los pilares más profundos de la organización: la confianza y la integridad emocional. La respuesta no consiste en eliminar la confianza, sino en redefinirla con inteligencia, procesos robustos y liderazgo consciente. La prevención del fraude moderno es, ante todo, un ejercicio de liderazgo humano, donde la alta dirección debe asumir la responsabilidad de cultivar una cultura resiliente, transparente y preparada para enfrentar los retos de la era digital. Solo así será posible justificar la necesidad de protocolos estrictos sin dañar la cultura organizacional, asegurando la supervivencia y el éxito corporativo en un entorno cada vez más incierto.

4. Marco Legal y Cumplimiento (México)

El fenómeno de los deepfakes en tiempo real plantea uno de los mayores desafíos contemporáneos para el derecho, el cumplimiento normativo y el gobierno corporativo. Aunque la tecnología avanza más rápido que la legislación, la inexistencia de una ley específica no implica ausencia de responsabilidad. Por el contrario, incrementa la carga de diligencia para la alta dirección.

Este capítulo tiene un objetivo claro a nivel CEO y Consejo: eliminar cualquier falsa sensación de vacío legal y dejar establecido que los deepfakes ya encajan, de manera directa y contundente, en múltiples figuras jurídicas vigentes en México.

Desde una perspectiva ejecutiva:

La pregunta legal no es si existe una ley de deepfakes, sino si la organización actuó con diligencia razonable frente a un riesgo conocido.

4.1 Fraude y usurpación de identidad

En el marco jurídico mexicano, los deepfakes en tiempo real encajan de forma natural en los delitos de fraude y usurpación de identidad, tanto a nivel federal como estatal.

Fraude

El fraude se configura cuando existe:

- Engaño
- Aprovechamiento del error
- Obtención de un beneficio indebido
- Perjuicio patrimonial o moral

El deepfake en tiempo real cumple plenamente con estos elementos al inducir a una persona a realizar actos que no habría ejecutado de conocer la verdad.

Ejemplos claros:

- Autorización de transferencias
- Firma o validación de contratos
- Liberación de información sensible

Usurpación de identidad

La suplantación de una persona real mediante imagen y voz clonadas constituye una forma avanzada de usurpación de identidad, aun cuando no exista contacto físico.

Punto clave para la alta dirección:

La identidad ejecutiva es un activo legal de la empresa; suplantarla genera responsabilidad penal y civil.

4.2 Delitos informáticos aplicables

Los deepfakes no operan en el vacío; dependen de medios digitales, lo que activa la aplicación de delitos informáticos ya tipificados.

Conductas relevantes

- Acceso ilícito a sistemas
- Manipulación de información
- Interceptación de comunicaciones
- Uso indebido de plataformas tecnológicas

Aunque el engaño sea psicológico, el medio es tecnológico, lo que habilita responsabilidades adicionales.

Desde compliance:

Un incidente de deepfake rara vez es un solo delito; suele ser una cadena de conductas ilícitas.

4.3 Protección de datos personales (LFPDPPP)

La Ley Federal de Protección de Datos Personales en Posesión de los Particulares (LFPDPPP) es plenamente aplicable a incidentes relacionados con deepfakes.

Datos comprometidos

- Datos de identificación
- Datos laborales
- Datos financieros
- Datos biométricos (imagen, voz)

Obligaciones del responsable

- Implementar medidas de seguridad administrativas, técnicas y físicas
- Prevenir accesos no autorizados
- Notificar vulneraciones relevantes

La imagen y la voz son datos personales, y su uso indebido sin consentimiento constituye una violación directa a la ley.

Para el Consejo:

No proteger la identidad digital de ejecutivos y colaboradores es una omisión en materia de datos personales.

4.4 Valor probatorio de video y voz

Históricamente, el video y la voz han sido considerados pruebas sólidas. En el contexto actual, su valor probatorio **ya no es automático**.

Riesgos probatorios

- Impugnación por posible manipulación
- Duda razonable sobre autenticidad
- Necesidad de peritajes especializados

Esto tiene implicaciones críticas:

- Contratos
- Autorizaciones
- Investigaciones internas
- Procedimientos judiciales

Mensaje clave:

Grabar no es probar; autenticar es probar.

Las organizaciones deben asumir que cualquier evidencia audiovisual puede ser cuestionada si no está respaldada por controles adicionales.

4.5 Responsabilidad corporativa por omisión

Uno de los mayores riesgos legales no proviene del ataque, sino de la falta de prevención.

Concepto de omisión

Existe responsabilidad cuando:

- El riesgo era conocido o previsible
- No se implementaron controles razonables
- El daño pudo mitigarse

Dado que los deepfakes ya son una amenaza documentada, la ignorancia deja de ser una defensa válida.

Esto impacta directamente a:

- Consejo de Administración
- CEO
- Directivos con funciones de control

Desde gobierno corporativo:

No actuar frente a un riesgo emergente conocido puede interpretarse como negligencia.

4.6 Buenas prácticas de defensa legal preventiva

La mejor defensa legal frente a un incidente de deepfake es demostrar diligencia debida.

Elementos clave

- Manuales y políticas formales
- Protocolos documentados de verificación
- Capacitación periódica
- Registro de decisiones y autorizaciones
- Auditorías internas
- Simulacros y pruebas

Beneficios legales

- Atenuación de responsabilidades
- Evidencia de buena fe
- Defensa ante reguladores
- Protección reputacional

Mensaje final al CEO:

Un manual no evita el ataque, pero protege a la organización y a su liderazgo cuando el ataque ocurre.

Conclusión ejecutiva

El marco legal mexicano ya ofrece herramientas suficientes para encuadrar y sancionar los deepfakes en tiempo real. La verdadera diferencia entre una empresa resiliente y una vulnerable radica en qué tan preparada estaba antes del incidente.

Este capítulo establece que la prevención no es opcional: es una obligación fiduciaria, legal y ética de la alta dirección.

Capítulo 5. Gobierno Corporativo y Responsabilidades en la Gestión del Riesgo de Deepfakes en Tiempo Real

Introducción

La gestión del riesgo de deepfakes en tiempo real ha dejado de ser un desafío meramente tecnológico para convertirse en un asunto estratégico central de gobierno corporativo. Su impacto trasciende lo operativo y afecta de lleno la integridad del liderazgo, la confianza de los grupos de interés, la toma de decisiones informada y la responsabilidad fiduciaria de quienes dirigen la organización. Un gobierno corporativo sólido no solo previene incidentes, sino que también garantiza rendición de cuentas y protección integral ante escenarios de crisis. La claridad en los roles, obligaciones y facultades de cada órgano y área es condición indispensable para transformar un riesgo complejo en uno gestionable y defendible.

5.1 Consejo de Administración

El Consejo de Administración representa la cúspide de la responsabilidad en materia de supervisión de riesgos, siendo el garante último de la protección del valor organizacional y del cumplimiento de las obligaciones fiduciarias hacia accionistas, colaboradores y sociedad. Su involucramiento en la gestión de deepfakes debe ser proactivo, estratégico y alineado con mejores prácticas internacionales.

Supervisión y liderazgo estratégico: El Consejo debe reconocer de manera formal el riesgo de deepfakes como un riesgo emergente de primer orden, incorporándolo a la matriz integral de riesgos y asegurando que su gestión reciba la misma prioridad que riesgos financieros, legales o de reputación.

Aprobación de políticas y marcos normativos: Es facultad exclusiva del Consejo aprobar el manual corporativo y las políticas relacionadas, estableciéndolas como documentos obligatorios y vinculantes para toda la organización. Debe exigir revisiones periódicas, asegurando su actualización frente a la evolución tecnológica y regulatoria.

Asignación y supervisión de recursos: El Consejo debe validar que existan recursos humanos, tecnológicos y presupuestales suficientes para la prevención, detección y respuesta ante deepfakes, monitoreando su aplicación real y la adecuada priorización frente a otros proyectos.

Monitoreo de incidentes y toma de decisiones extraordinarias: Ante incidentes de alto impacto, el Consejo supervisa la actuación del CEO, valida decisiones que puedan comprometer la reputación, finanzas o continuidad del negocio, y participa en la definición de estrategias de comunicación y remediación.

Adopción de mejores prácticas internacionales: El Consejo debe mantenerse informado sobre estándares globales en gestión de riesgos digitales, promoviendo benchmarking y la adopción de marcos de referencia reconocidos (ej. COSO, ISO 31000, NIST).

Rendición de cuentas y mejora continua: Es responsable de solicitar reportes periódicos sobre el estado de madurez, resultados de auditorías, simulacros y métricas clave, asegurando la retroalimentación y mejora constante del sistema de gestión de riesgos.

Mensaje rector: El Consejo no ejecuta controles, pero responde por no haberlos exigido. La omisión o ambigüedad en sus mandatos es, en sí misma, un riesgo estratégico.

5.2 CEO y Alta Dirección

El CEO y su equipo directivo asumen la responsabilidad indelegable de implementar y ejecutar efectivamente la estrategia de prevención y respuesta frente a deepfakes. Su liderazgo visible y su compromiso con la cultura de verificación marcan la diferencia entre una organización resiliente y una vulnerable.

Establecimiento del "tono desde la cima": El CEO debe comunicar de manera clara y constante que la verificación no es una muestra de desconfianza, sino una práctica esencial de lealtad institucional y protección colectiva.

Respaldar y cumplir personalmente los protocolos: El liderazgo debe ser ejemplar, sometiéndose a los mismos controles y verificaciones que exige a sus equipos, y defendiendo públicamente los procedimientos, incluso ante incomodidades o presiones.

Comunicación institucional en crisis: Ante incidentes, el CEO lidera la comunicación interna y externa, asegurando mensajes unificados, oportunos y alineados con la estrategia legal y reputacional.

Prevención de decisiones basadas en voz, imagen o urgencia emocional: El CEO debe garantizar que ninguna decisión crítica dependa exclusivamente de estos factores, promoviendo sistemas de control dual y validación independiente.

Traducción estratégica a lo operativo: La Alta Dirección debe convertir las políticas en procedimientos específicos para cada área, identificar procesos vulnerables, capacitar a su personal y escalar cualquier sospecha de inmediato.

Gestión de incidentes y lecciones aprendidas: Tras un evento, el CEO debe liderar la evaluación integral, asegurando la identificación de causas raíz, la implementación de mejoras y la retroalimentación a todos los niveles.

Principio rector: La jerarquía no sustituye la verificación; la obliga. El liderazgo se demuestra en la congruencia y en la rendición de cuentas.

5.3 Seguridad Corporativa

Seguridad Corporativa es el integrador transversal del riesgo de deepfakes, responsable de articular personas, procesos y tecnología bajo un enfoque de protección integral y respuesta coordinada.

Diseño, mantenimiento y actualización de protocolos: Debe desarrollar, revisar y perfeccionar los procedimientos anti-deepfake, asegurando su alineación con las mejores prácticas y su adaptación dinámica ante nuevas amenazas.

Coordinación de la detección, respuesta y recuperación: Es responsable de activar mecanismos de alerta temprana, liderar la contención de incidentes, gestionar la recuperación operativa y documentar todas las acciones.

Definición de criterios claros de activación: Debe establecer umbrales y procedimientos objetivos para la activación de protocolos, evitando zonas grises y asegurando la actuación oportuna sin demoras injustificadas.

Capacitación y sensibilización: Coordina entrenamientos periódicos para personal clave, simulacros y ejercicios de concientización, midiendo el nivel de preparación y respuesta de los equipos.

Integración con planes de crisis y continuidad: Asegura que el riesgo de deepfakes esté contemplado en los planes de continuidad del negocio y gestión de crisis, con roles y responsabilidades definidos.

Autoridad en incidentes: Tiene la facultad para detener decisiones críticas, aislar procesos comprometidos, preservar evidencia digital y coordinar con áreas legales, TI y comunicación.

Nota: Seguridad Corporativa no autoriza pagos ni contratos, pero cuenta con autoridad para frenar cualquier decisión bajo sospecha razonable.

5.4 Auditoría Interna

Auditoría Interna es el órgano independiente encargado de garantizar la efectividad, consistencia y auditabilidad de los controles implementados frente al riesgo de deepfakes.

Evaluación integral de controles: Revisa el diseño, implementación y funcionamiento real de los controles, más allá de la existencia documental, identificando brechas y áreas de mejora.

Revisiones periódicas y auditorías temáticas: Realiza auditorías programadas y aleatorias, incluyendo simulacros, análisis de incidentes y revisiones de cumplimiento específico.

Análisis forense de incidentes y casi-incidentes: Documenta, analiza y reporta hallazgos sobre eventos reales o potenciales, extrayendo lecciones para fortalecer el sistema de gestión.

Reporte directo al Consejo: Presenta resultados, hallazgos y recomendaciones directamente al Consejo de Administración, asegurando independencia y transparencia.

Fomento de la mejora continua: Propone ajustes estructurales, promueve la adopción de mejores prácticas y da seguimiento al cierre de hallazgos y recomendaciones.

Principio rector: Auditoría evalúa sistemas y procesos, no personas. El enfoque es preventivo y correctivo, nunca punitivo.

5.5 Cumplimiento / Legal

Las áreas de Cumplimiento y Legal convierten el riesgo de deepfakes en obligaciones normativas claras y defensas jurídicas sólidas para la organización y sus líderes.

Alineación con la legislación mexicana vigente: Analizan y adaptan los protocolos para asegurar su concordancia con marcos normativos nacionales, internacionales y sectoriales, incluyendo protección de datos personales, ciberseguridad y responsabilidad corporativa.

Evaluación de riesgos legales y de responsabilidad por omisión: Identifican posibles escenarios de responsabilidad penal, civil, administrativa o reputacional por falta de previsión, acción o documentación.

Definición de criterios de notificación y reporte: Establecen cuándo, cómo y ante qué autoridades se deben notificar incidentes, asegurando trazabilidad y cumplimiento de plazos legales.

Preparación de estrategias legales y de comunicación post-incidente: Colaboran en la definición de mensajes clave, respuestas a medios y estrategias de defensa ante posibles investigaciones o litigios.

Aseguramiento de la trazabilidad documental: Supervisan que cada acción, decisión y comunicación esté debidamente registrada y resguardada para efectos de defensa jurídica.

Capacitación legal: Desarrollan programas de sensibilización y formación en materia de cumplimiento, ética y prevención de riesgos legales asociados a deepfakes.

Mensaje clave: Lo que no está documentado, no existe jurídicamente. La defensa legal se construye desde la prevención y la evidencia.

5.6 Recursos Humanos

Recursos Humanos es responsable de gestionar el impacto humano y cultural del riesgo de deepfakes, promoviendo una cultura de prevención, confianza y resiliencia.

Programas de capacitación y sensibilización continua: Desarrolla e implementa entrenamientos periódicos sobre riesgos de suplantación, procedimientos de reporte y manejo de incidentes.

Inclusión del riesgo en procesos de onboarding y offboarding: Asegura que todo nuevo colaborador reciba formación sobre deepfakes y que, al egreso, se refuercen medidas de protección de información y acceso.

Protección reforzada de datos personales y biométricos: Implementa controles adicionales para el resguardo de información sensible, minimizando la exposición a ataques de suplantación.

Gestión emocional y contención post-incidente: Proporciona apoyo psicológico y asesoría a colaboradores afectados, evitando la estigmatización o cultura de culpa.

Fomento de una cultura organizacional de confianza: Promueve el reporte inmediato de sospechas y la colaboración transversal, sin temor a represalias o sanciones indebidas.

Nota: RH es custodio de la confianza interna y motor de la resiliencia organizacional.

5.7 TI / Ciberseguridad

Las áreas de TI y Ciberseguridad proporcionan el soporte tecnológico habilitante para la prevención, detección y respuesta ante deepfakes, en coordinación estrecha con Seguridad Corporativa y Legal.

Implementación de plataformas y herramientas seguras: Evalúan, seleccionan e implementan tecnologías de autenticación robusta, análisis de voz e imagen, y mecanismos de detección automatizada de deepfakes.

Monitoreo de patrones anómalos y amenazas emergentes: Mantienen sistemas de monitoreo en tiempo real, identificando comportamientos inusuales y gestionando alertas de manera prioritaria.

Soporte en análisis forense: Proveen capacidades técnicas para la investigación de incidentes, preservando evidencia digital y colaborando en la reconstrucción de eventos.

Evaluación y gestión de riesgos de proveedores tecnológicos: Auditan la seguridad de soluciones externas, asegurando el cumplimiento de políticas internas y la actualización constante frente a nuevas vulnerabilidades.

Mantenimiento de la resiliencia operativa: Garantizan la continuidad y recuperación de sistemas críticos, minimizando el impacto de incidentes y facilitando la reanudación segura de operaciones.

Capacitación técnica especializada: Forman a usuarios clave en el uso seguro de herramientas, detección de señales de alerta y procedimientos de escalamiento.

Principio rector: La tecnología apoya la gestión del riesgo, pero no reemplaza el criterio humano ni la supervisión directiva.

5.8 Proveedores y Terceros

Los proveedores y terceros amplían la superficie de riesgo y, por tanto, deben ser gestionados bajo criterios estrictos de seguridad, cumplimiento y colaboración.

Obligaciones contractuales explícitas: Todos los contratos deben incluir cláusulas claras sobre políticas anti-deepfake, manejo de incidentes y consecuencias ante incumplimientos.

Limitación de accesos a lo estrictamente necesario: Se debe aplicar el principio de mínimo privilegio, restringiendo accesos y permisos solo a lo indispensable para la prestación del servicio.

Notificación inmediata de incidentes: Los proveedores están obligados a reportar cualquier sospecha o incidente de deepfake que pueda afectar a la organización, facilitando la respuesta coordinada.

Aceptación de auditorías y revisiones: Deben permitir auditorías periódicas, revisiones técnicas y la implementación de mejoras derivadas de hallazgos conjuntos.

Evaluación y homologación de proveedores: La selección debe considerar criterios de madurez en la gestión de riesgos digitales, historial de cumplimiento y capacidad de respuesta ante incidentes.

Principio rector: La organización es tan segura como su proveedor más débil. La gestión proactiva de terceros es parte integral de la resiliencia corporativa.

Conclusión Ejecutiva

Un gobierno corporativo robusto y claramente estructurado es el escudo más efectivo frente a la amenaza de deepfakes en tiempo real. La definición exhaustiva, explícita y no negociable de roles, obligaciones y facultades elimina zonas grises, previene la parálisis y fortalece la capacidad de respuesta ante crisis. La prevención es una obligación fiduciaria, legal y ética de la alta dirección; la rendición de cuentas, un pilar irrenunciable de la confianza organizacional. Solo la acción coordinada, informada y transparente protegerá a la empresa, a sus líderes y a todos sus grupos de interés frente a este desafío emergente.

Capítulo 6. Clasificación de Riesgos por Área

La amenaza de deepfakes no afecta de manera uniforme a todas las áreas de la organización. Su peligrosidad varía según el tipo de proceso, el momento de la toma de decisiones y el nivel de autoridad involucrado. Esta clasificación exhaustiva permite identificar los puntos de mayor exposición, asignar responsabilidades precisas y diseñar controles específicos, transformando el concepto abstracto del riesgo en un mapa operativo concreto y accionable.

6.1 Finanzas y Tesorería

El área financiera y de tesorería se posiciona como el blanco principal de los ataques con deepfake debido a la naturaleza crítica de sus operaciones y al valor económico inmediato de las transacciones que gestiona. Los riesgos no solo involucran la transferencia de

fondos, sino también la gestión de cuentas, validación de pagos y la interacción con entidades externas (bancos, auditores, socios).

Escenarios frecuentes incluyen la recepción de instrucciones falsas de transferencia urgente, supuestamente emitidas por altos directivos, videollamadas manipuladas para solicitar la ruptura de controles bajo el argumento de confidencialidad, suplantación de entidades externas y confirmaciones de pago mediante voz clonada. Estos ataques aprovechan la urgencia y jerarquía para generar presión y acelerar decisiones sin validación suficiente.

El impacto puede ser devastador: pérdidas económicas directas, operaciones irreversibles, incumplimientos contractuales y afectación a la confianza de inversionistas. Para mitigar estos riesgos, se deben implementar controles estrictos, como la prohibición de validar cualquier instrucción financiera crítica solo por voz o video, la separación de funciones y la doble validación fuera del canal original, además de establecer ventanas de enfriamiento para decisiones urgentes. La certeza debe prevalecer sobre la velocidad, sin excepciones.

6.2 Compras y Contrataciones

El área de compras y contrataciones enfrenta riesgos derivados de la manipulación de flujos de pago y la introducción de proveedores o terceros fraudulentos. El uso de deepfakes puede facilitar el cambio de cuentas bancarias mediante videollamadas falsas, la aprobación acelerada de contratos por instrucciones supuestamente directivas y la manipulación emocional que apela al “riesgo de perder al proveedor”.

El impacto se traduce en fraudes recurrentes de bajo monto, difíciles de detectar por su dispersión y frecuencia, comprometiendo la cadena de suministro y exponiendo a la organización a riesgos de corrupción y colusión. Como controles esenciales, se debe establecer la verificación contractual multicanal, mantener un registro formal de cambios sensibles y prohibir cualquier modificación contractual por canales no institucionales. Ante la urgencia de un proveedor, la alerta debe activarse antes que la prioridad.

6.3 Recursos Humanos

Recursos Humanos es particularmente vulnerable debido al manejo de información confidencial de empleados, decisiones de nómina y la gestión de relaciones laborales. Los deepfakes pueden materializarse en órdenes falsas de despido o contratación, solicitudes de información delicada y manipulación de conflictos internos.

El impacto va más allá de lo legal, afectando la cultura organizacional y exponiendo a la empresa a violaciones a la Ley Federal de Protección de Datos Personales en Posesión de los Particulares (LFPDPPP), demandas laborales y daño reputacional interno. Los controles clave incluyen la prohibición de tomar decisiones laborales basadas únicamente en llamadas, la validación documental obligatoria y la capacitación psicológica para identificar autoridad falsa. La dignidad y seguridad del empleado no pueden depender de una videollamada o instrucción no verificada.

6.4 Seguridad y Crisis

El área de seguridad y gestión de crisis enfrenta riesgos de manipulación directa en situaciones de alto estrés. Los deepfakes pueden generar órdenes falsas de evacuación, manipulación de incidentes de seguridad y confusión en centros de mando, lo que puede derivar en riesgo físico, caos organizacional y pérdida de control operativo.

La respuesta debe estar sustentada en la autoridad clara y jerarquía validada, protocolos de verificación fuera del canal habitual y la realización periódica de simulacros para fortalecer la preparación y reducir la vulnerabilidad ante eventos simulados o agravados por deepfakes. La gestión de crisis exige rigor y control, nunca improvisación.

6.5 Comunicación Corporativa

La comunicación corporativa es el amplificador reputacional del deepfake. Videos falsos del CEO, comunicados internos apócrifos y crisis mediáticas fabricadas pueden dañar la imagen institucional, erosionar la credibilidad y afectar la percepción de mercados y reguladores.

Los controles esenciales incluyen la vocería única, la certificación de canales oficiales y el monitoreo activo de medios y redes sociales para detectar y neutralizar comunicaciones falsas. Un silencio desordenado puede ser más perjudicial que una respuesta lenta pero verificada; la coordinación y transparencia son fundamentales en la gestión reputacional frente a deepfakes.

6.6 Jurídico y Compliance

El área jurídica y de cumplimiento se enfrenta a la responsabilidad directa por omisión o manejo inadecuado de incidentes relacionados con deepfakes. Los riesgos incluyen la falta de denuncia oportuna, la gestión incorrecta de evidencia y el incumplimiento normativo posterior al incidente.

El impacto puede derivar en sanciones regulatorias, litigios prolongados y responsabilidad personal de los directivos. Los controles indispensables comprenden protocolos de preservación de evidencia, coordinación inmediata con seguridad y documentación exhaustiva del incidente. Un evento mal documentado se convierte en un problema legal asegurado; la gestión debe ser proactiva y rigurosa.

Clasificar los riesgos por área funcional permite a la organización transformar la amenaza de los deepfakes en un mapa de riesgos concreto, priorizar recursos, diseñar controles específicos y asegurar que cada área entienda su exposición real y responsabilidad directa.

Capítulo 7. Controles Preventivos (Antes del Incidente)

La prevención es el único momento en el ciclo de riesgo donde la organización tiene control total. Una vez que un deepfake en tiempo real está activo, el margen de error se reduce drásticamente y el costo de una mala decisión se multiplica. Los controles preventivos son reglas de gobierno corporativo, no meras recomendaciones tecnológicas,

y deben ser adoptados de manera estricta por los ejecutivos, mandos críticos y áreas sensibles.

7.1 Principios de verificación de identidad

La verificación de identidad debe ser rigurosa y nunca asumida por apariencia o familiaridad. El principio de cero confianza por imagen o voz implica que ninguna instrucción crítica puede considerarse válida solo porque “parece” o “suena” como alguien conocido. La separación entre identidad y autoridad obliga a verificar, incluso a quienes ostentan cargos altos. Toda instrucción sensible debe confirmarse por un canal distinto al que la originó, garantizando trazabilidad total y documentación auditable.

El derecho y obligación de dudar debe ser institucional; cualquier colaborador debe pausar y escalar si algo no es verificable. La desconfianza bien diseñada protege, mientras que la confianza ciega expone a la organización.

7.2 Prohibiciones críticas (qué NO se debe hacer)

Las acciones prohibidas son claras y no admiten excepciones. Está terminantemente prohibido autorizar pagos, transferencias o cambios bancarios por llamada, videollamada o mensaje; ejecutar instrucciones “confidenciales” que pidan romper controles; usar canales personales como WhatsApp, Telegram o llamadas privadas para decisiones críticas; compartir códigos, contraseñas o accesos por voz o video; y actuar bajo presión emocional inducida (urgencia, miedo, secreto).

Si una instrucción pide saltarse controles, es una señal inequívoca de ataque, no de liderazgo. Estas prohibiciones deben estar comunicadas y reforzadas de manera sistemática en toda la organización.

7.3 Autorizaciones de alto impacto

Las autorizaciones de alto impacto comprenden movimientos de recursos financieros, modificaciones de datos sensibles, decisiones laborales críticas, comunicación externa relevante y activación de protocolos de crisis. Deben estar sujetas a doble o triple validación independiente, con separación estricta de funciones, registro formal previo a la ejecución y tiempo mínimo de reflexión (cooling-off) para evitar decisiones precipitadas.

La urgencia nunca justifica la omisión de validaciones; cada autorización debe ser analizada y confirmada en etapas, sin excepción.

7.4 Verificación multicanal

Confirmar identidad y autorización mediante canales diferentes, independientes y controlados es fundamental. Los canales válidos incluyen sistemas corporativos autenticados, correos institucionales con firma digital, plataformas internas seguras y contacto presencial. Quedan excluidos las redes sociales, mensajería instantánea no corporativa y llamadas no grabadas. Si no puede verificarse fuera del canal, no puede ejecutarse. Esta regla debe ser aplicada con rigor en todos los niveles de la organización.

7.5 Frases o códigos dinámicos

Agregar una capa humana de autenticación mediante frases o códigos dinámicos dificulta la replicación en tiempo real por parte de atacantes. Estos códigos cambian periódicamente, son conocidos solo por personas autorizadas, nunca son reutilizables ni relacionados con datos personales y se activan únicamente ante solicitudes críticas, verificadas fuera del canal inicial y registradas como evento de control.

Un código estático equivale a una contraseña disfrazada y no debe emplearse como mecanismo de autenticación en procesos críticos.

7.6 Gestión de accesos y privilegios

La gestión de accesos debe regirse por el principio de mínimo privilegio, con accesos temporales y justificados, revisión periódica obligatoria y revocación inmediata ante incidentes o cambios en la estructura organizacional. Los riesgos comunes incluyen accesos heredados, cuentas compartidas y falta de segregación, lo que abre puertas innecesarias al engaño y la manipulación.

Cada acceso innecesario representa una vulnerabilidad potencial; la administración de privilegios debe ser proactiva y dinámica, adaptándose a la evolución de riesgos y necesidades operativas.

7.7 Capacitación y concientización

Asimismo, fortalecer la resiliencia organizacional requiere integrar la capacitación constante con procesos de mejora continua, donde cada experiencia, simulacro o incidente real sirva para refinar protocolos y anticipar nuevas tácticas de manipulación. La actualización periódica de contenidos, el intercambio de aprendizajes entre áreas y la creación de espacios seguros para reportar sospechas sin represalias consolidan una defensa colectiva adaptable. Solo así, conjugando el criterio humano entrenado con la robustez de los controles operativos, es posible convertir el riesgo de los deepfakes en una variable gestionable, permitiendo a la organización responder con agilidad y cohesión ante amenazas emergentes.

En conclusión, la implementación rigurosa y detallada de la clasificación de riesgos por área y los controles preventivos transforma la amenaza de los deepfakes en un riesgo gobernado. Antes del incidente, la organización decide con método y responsabilidad; después, solo queda mitigar consecuencias. Adoptar estos controles no es una opción, sino una obligación fiduciaria y ética para el liderazgo corporativo, asegurando la seguridad, resiliencia y confianza institucional ante un entorno cada vez más complejo y desafiante.

Capítulo 8. Detección de Deepfake en Tiempo Real

Introducción: Importancia de la Detección en Tiempo Real

En el contexto actual, donde la sofisticación de los ataques digitales evoluciona a la par de los controles, la detección en tiempo real de deepfakes representa la línea divisoria entre una organización resiliente y una vulnerable. La capacidad de identificar y actuar oportunamente ante intentos de manipulación sintética no solo minimiza riesgos operativos y reputacionales, sino que también fortalece la confianza institucional y la capacidad de respuesta ante amenazas emergentes. Las organizaciones que implementan mecanismos robustos de detección logran transformar el deepfake de una amenaza impredecible en un riesgo gestionado y controlable.

Es fundamental comprender que la detección efectiva no reside únicamente en el uso de tecnología avanzada, sino en el desarrollo del criterio humano, la observación sistemática y el empoderamiento de los equipos para detener interacciones sospechosas sin temor a represalias ni dudas sobre su autoridad.

Principio Rector: Detectar No es Probar

Uno de los errores más comunes en la gestión de incidentes de deepfake es asumir que la detección requiere prueba absoluta. Sin embargo, en el marco de la seguridad corporativa, detectar significa identificar anomalías suficientes para suspender o detener un proceso antes de que se consume un daño. Probar implica un análisis forense posterior, generalmente fuera del alcance y responsabilidad de quienes toman decisiones en tiempo real.

Este principio rector establece que la prevención se basa en la capacidad de reconocer señales atípicas, aun cuando sean sutiles o ambiguas, y actuar de inmediato. El objetivo no es demostrar la existencia de un deepfake en el momento, sino evitar que la organización se vea comprometida por no haber intervenido a tiempo.

8.1 Indicadores Técnicos

Los indicadores técnicos son manifestaciones observables en los canales de comunicación audiovisual que, individualmente, pueden pasar desapercibidas o atribuirse a fallas técnicas comunes; sin embargo, al acumularse, revelan la presencia potencial de manipulación sintética. Reconocer estos indicadores requiere entrenamiento específico y atención minuciosa a los detalles.

Indicadores en Video:

Micro-desfase entre audio y movimiento labial:

Se detecta cuando el sonido de las palabras no coincide perfectamente con el movimiento de los labios. Un retraso de fracciones de segundo puede indicar que el video ha sido manipulado, especialmente si el desfase persiste o se incrementa durante la conversación.

Parpadeo irregular o inexistente:

Los deepfakes suelen presentar parpadeos poco naturales: demasiado espaciados, repetitivos o, en algunos casos, completamente ausentes. El parpadeo humano es espontáneo y variable; su alteración debe considerarse sospechosa.

Expresiones faciales rígidas o desincronizadas:

Las expresiones pueden parecer forzadas, mantener una rigidez inusual o no corresponder con el tono emocional del mensaje. Esto ocurre porque los algoritmos aún tienen dificultades para emular la sutileza y fluidez de los gestos humanos.

Iluminación inconsistente en rostro vs. entorno:

Observa si la iluminación del rostro no concuerda con la del fondo o el entorno, generando sombras o brillos artificiales. Este desajuste suele ser resultado de la superposición de rostros sintéticos sobre fondos reales.

Bordes faciales borrosos o inestables:

Los contornos del rostro pueden aparecer difuminados, vibrar o presentar cambios abruptos, especialmente cuando la persona gira la cabeza o realiza movimientos bruscos.

Indicadores en Audio:

Voz excesivamente limpia o constante:

Una voz sin variaciones naturales, carente de ruido de fondo o con un timbre demasiado uniforme, puede ser síntoma de síntesis artificial.

Falta de respiración natural:

La ausencia de sonidos de respiración, inhalaciones o exhalaciones, es un indicador relevante, ya que los generadores de voz suelen omitir estos detalles.

Cambios súbitos de entonación:

Variaciones abruptas en el tono, volumen o ritmo que no corresponden al contexto emocional son señales de edición o generación sintética.

Respuestas con latencia artificial:

Pequeños retrasos en las respuestas, especialmente si se repiten de forma sistemática, pueden indicar procesamiento adicional típico de sistemas automatizados.

Advertencia ejecutiva: Un deepfake no necesita verse mal; muchas veces, su perfección excesiva es precisamente la señal de alerta. Si una imagen o voz parecen “demasiado perfectas”, debe activarse el protocolo de verificación.

8.2 Indicadores Conductuales

Los indicadores conductuales se refieren a patrones de comportamiento y comunicación que, aunque difíciles de falsificar, pueden ser reproducidos parcialmente por atacantes

bien informados. Por ello, el análisis debe ser holístico y considerar tanto el contexto como la dinámica de la interacción.

Evita preguntas inesperadas:

Cuando el interlocutor esquiva o responde de forma ambigua a preguntas fuera del guion, especialmente aquellas que requieren conocimiento contextual o reciente, es motivo de sospecha. Los deepfakes suelen estar programados para responder solo a temas previstos.

Redirige constantemente a la urgencia:

El atacante insiste en la inmediatez de la acción, minimizando el tiempo para reflexionar o consultar. Frases como “esto no puede esperar” o “es crítico que lo hagas ahora” son comunes.

Repite frases clave:

El uso reiterado de palabras o instrucciones específicas busca reforzar la manipulación y evitar que la víctima cuestione el mensaje.

No tolera pausas ni verificaciones:

Ante cualquier intento de detener la interacción para validar datos o consultar con terceros, el deepfake presiona para continuar, argumentando falta de tiempo o confidencialidad.

Solicita secreto absoluto:

Se exige que la acción o información no sea compartida con nadie más, bajo pretextos de seguridad, confidencialidad o riesgo inminente.

Patrón típico: Cuando se combina autoridad percibida (posición jerárquica), urgencia extrema y aislamiento de la víctima, existe una alta probabilidad de que se trate de un intento de ataque en curso.

8.3 Señales de Alerta Psicológica

Los deepfakes no solo manipulan la percepción visual y auditiva, sino que explotan respuestas automáticas del cerebro humano para incrementar la probabilidad de éxito. Identificar síntomas internos es tan importante como reconocer indicadores externos.

Aumento súbito de estrés:

Sensación de presión o ansiedad sin causa aparente, generalmente provocada por la urgencia o el tono de la interacción.

Sensación de “no pensar demasiado”:

Impulso a actuar rápidamente, evitando la reflexión o el análisis profundo, motivado por la insistencia del interlocutor.

Miedo a desobedecer:

Temor a consecuencias negativas por no cumplir la instrucción, especialmente si proviene de una supuesta figura de autoridad.

Confusión cognitiva:

Dificultad para ordenar ideas, recordar procedimientos o evaluar la lógica de la solicitud, resultado de la sobrecarga emocional inducida.

Estrategias de defensa: Reconocer estos síntomas y verbalizarlos ante el equipo de seguridad permite activar protocolos de protección antes de que la manipulación sea irreversible.

Mensaje clave: Si te sientes presionado a no pensar o a actuar sin consultar, ya estás bajo ataque. La autopercepción es una herramienta crítica de defensa.

8.4 Latencia, Sincronía y Patrones Anómalos

Las anomalías en la interacción, aunque a veces justificables por problemas técnicos, pueden ser indicio de manipulación cuando se presentan de manera persistente o junto a otros indicadores.

Latencia:

Respuestas ligeramente retrasadas:

Un retraso constante de uno o dos segundos en la contestación puede deberse al procesamiento de sistemas automatizados o a la edición en tiempo real.

Silencios artificiales antes de contestar:

Pausas poco naturales, especialmente cuando no hay ruido ambiental, denotan que la respuesta está siendo generada o seleccionada.

Sincronía:

Movimientos que no coinciden con el discurso:

Gestos, movimientos de cabeza o manos que no corresponden al énfasis verbal, o que parecen desfasados respecto a lo que se dice.

Gestos repetitivos:

Repetición de los mismos movimientos (asentir, sonreír, mirar hacia un lado) en intervalos regulares, resultado de la programación limitada del deepfake.

Patrones:

Conversaciones lineales sin espontaneidad:

Intercambios demasiado estructurados, sin cambios de tema, bromas, comentarios personales ni respuestas improvisadas.

Incapacidad de improvisar detalles contextuales:

Dificultad para responder a preguntas sobre eventos recientes, situaciones fuera del guion o detalles específicos de la organización.

Principio práctico: La imperfección humana —errores al hablar, titubeos, cambios de tono, interrupciones— es un factor de autenticidad. La ausencia de estas imperfecciones debe disparar las alertas de verificación.

8.5 Reporte Inmediato de Sospecha

El éxito de la detección en tiempo real depende de la capacidad de reportar sospechas sin demora, sin justificar la decisión y bajo un protocolo claro y protegido.

Regla de oro: La sospecha se reporta, no se justifica. No se requiere certeza ni evidencia concluyente para activar el protocolo; basta con identificar una anomalía significativa.

Protocolo ampliado:

Detener la interacción: Suspende la conversación de inmediato, utilizando frases institucionales como “Debo validar esta solicitud conforme a protocolo”.

No confrontar al interlocutor: Evita discutir, acusar o revelar sospechas al posible atacante; esto previene represalias o escalamiento del ataque.

Notificar a Seguridad Corporativa: Utiliza los canales oficiales, previamente definidos y autenticados, para informar del incidente.

Preservar evidencia: Guarda registros de la comunicación (video, audio, mensajes), así como notas de las anomalías detectadas, para facilitar el análisis posterior.

Recibir orientación: Sigue las instrucciones del área de Seguridad para determinar los pasos siguientes y evitar la diseminación del ataque.

Protección institucional:

Prohibición expresa de represalias: Ningún colaborador puede ser sancionado por reportar una sospecha, aun si resulta ser un falso positivo.

Reconocimiento a la detección temprana: Se debe valorar y premiar la identificación oportuna de riesgos, fortaleciendo la cultura de prevención y participación.

Mensaje final: Un falso positivo implica una breve interrupción; un falso negativo puede ocasionar pérdidas millonarias y daños irreparables. La rapidez y claridad en el reporte son la mejor defensa.

Conclusión Ejecutiva

La detección de deepfakes en tiempo real no exige que cada colaborador sea un experto forense, sino que desarrolle una duda informada, respaldada por la organización y sustentada en protocolos claros. El verdadero diferenciador entre caer o resistir ante un

deepfake es la disposición a cuestionar lo inusual y la confianza para actuar con respaldo institucional.

El liderazgo corporativo debe garantizar que todo el personal cuente con herramientas, capacitación y protección necesarias para ejercer este criterio. Prevenir es siempre más eficiente y menos costoso que reaccionar. Así, la detección proactiva y el reporte inmediato consolidan una cultura organizacional resiliente, capaz de adaptarse y responder ante amenazas cada vez más sofisticadas.

En definitiva, la diferencia entre una crisis y una anécdota radica en la oportunidad de reacción y la calidad de la decisión en el momento clave. Que la duda razonable, el conocimiento técnico y el respaldo institucional sean siempre la primera línea de defensa.

Capítulo 9. Protocolo de Respuesta Inmediata

Este capítulo define qué hacer cuando la prevención ya no fue suficiente y la detección ha levantado una sospecha real. Aquí no hay espacio para improvisación. Los primeros minutos determinan si el impacto será contenido o amplificado.

Un protocolo de respuesta inmediata existe para proteger a la organización, a sus personas y a sus líderes, no para buscar culpables. Su correcta activación es una señal de madurez, no de fracaso.

Principio rector:

En una crisis de identidad, la velocidad sin método es el mayor enemigo.

9.1 Activación del protocolo

La activación del protocolo de respuesta inmediata es el primer y más crucial paso cuando surge una sospecha de deepfake o manipulación de identidad. No se debe esperar a tener certeza absoluta, ya que el tiempo es fundamental: cada minuto de indecisión aumenta el riesgo de daño. La sospecha razonable, aunque no tenga pruebas técnicas contundentes, es suficiente para actuar, pues los atacantes suelen aprovechar los vacíos y demoras para consolidar el fraude.

Entre los disparadores clave para activar el protocolo están: el recibo de instrucciones críticas por voz o video que no pueden ser verificadas de manera confiable, la presencia de presión inusual o insistente para que se salten controles o procedimientos normales, la acumulación de señales técnicas (como anomalías en la calidad de la llamada o el video), comportamentales (cambios en la forma de comunicarse del supuesto emisor) o psicológicas (tono de urgencia, amenazas veladas, manipulación emocional), así como las dudas fundadas de cualquier colaborador, independientemente de su rango.

La activación del protocolo está diseñada para empoderar a todos: cualquier colaborador expuesto a una instrucción crítica, sin importar su jerarquía, está facultado para iniciar el

proceso. Mandos medios, directivos y el área de Seguridad Corporativa también pueden hacerlo. No se requiere autorización previa, precisamente porque la rapidez es esencial. Es fundamental entender que activar el protocolo no es un acto de desobediencia, sino de protección institucional y personal; es una muestra de madurez y cultura preventiva.

9.2 Aislamiento de la decisión

El aislamiento busca cortar de raíz cualquier decisión que podría haber sido manipulada. Esto implica suspender inmediatamente la ejecución de la instrucción sospechosa, lo que puede incluir detener transferencias de fondos, bloquear accesos privilegiados o pausar la comunicación con el presunto emisor. Esta suspensión es temporal, pero vital para evitar que el daño se propague, se normalice o se vuelva irreversible.

No basta con detener el primer paso; también es importante bloquear cualquier intento de reenvío, copia o confirmación de la instrucción, así como minimizar la interacción con el posible atacante. La comunicación debe ser neutral, sin confrontaciones ni explicaciones innecesarias al supuesto emisor, ya que cualquier información adicional podría ser utilizada en contra de la organización.

El principio operativo es claro: mientras exista duda sobre la identidad del emisor, ninguna decisión crítica puede continuar. El aislamiento se realiza sin la necesidad de explicar ni justificar al presunto emisor los motivos, lo que protege tanto al colaborador como a la organización de represalias o manipulaciones adicionales.

9.3 Verificación por canal seguro

La verificación es el proceso que permite confirmar la autenticidad de la instrucción recibida, pero jamás debe realizarse por el mismo canal donde surgió la sospecha. Los atacantes pueden controlar ese canal o replicar comportamientos creíbles, por lo que se deben utilizar únicamente medios corporativos certificados, como sistemas internos autenticados, líneas telefónicas oficiales o plataformas de mensajería protegidas.

Los métodos válidos de verificación incluyen: confirmar la instrucción mediante registros en sistemas internos con autenticación robusta, realizar una llamada de retorno a números oficiales previamente registrados en la organización, solicitar una confirmación escrita con firma digital o realizar una validación cruzada con al menos dos responsables independientes del área.

En situaciones donde es posible, la verificación presencial o la revisión documental física pueden agregar una capa extra de certeza. Es importante destacar que, aunque un deepfake puede sostener una conversación convincente, no puede atravesar un ecosistema de controles bien diseñados y ejecutados. El objetivo es que la verificación sea robusta, no burocrática, y que la seguridad nunca dependa de un solo canal o persona.

9.4 Escalamiento jerárquico

El escalamiento jerárquico garantiza que una situación que podría ser crítica no quede en manos de una sola persona, sino que pase rápidamente a niveles institucionales donde

existen mayores recursos, experiencia y autoridad para gestionar el evento. La cadena de escalamiento mínima comienza con Seguridad Corporativa, sigue con la dirección del área afectada, pasa por los departamentos de Cumplimiento o Legal y, dependiendo del impacto, llega hasta el CEO y, en casos extremos, al Consejo de Administración.

El proceso de escalamiento debe ser ágil: la prioridad es informar los hechos tal como sucedieron, sin interpretaciones ni juicios personales, para que las decisiones se basen en información objetiva y centralizada. La gestión de crisis requiere que la información fluya de manera controlada y que las decisiones se tomen en un solo punto, evitando la dispersión y el caos. Una crisis distribuida, donde cada quien actúa por su cuenta, se vuelve incontrolable, mientras que una crisis centralizada permite respuestas coordinadas y efectivas.

9.5 Registro y preservación de evidencia

La documentación minuciosa es fundamental para la defensa legal, la investigación forense y la mejora de los controles internos. Cada grabación de audio o video, registro de llamadas o mensajes, metadato asociado y testimonio inmediato debe ser preservado íntegramente. No se deben editar, manipular ni alterar los archivos originales bajo ninguna circunstancia, ya que esto podría invalidar la evidencia en procedimientos legales o administrativos.

Es esencial asegurar la integridad y la cadena de custodia de toda la evidencia, lo que implica almacenarla en repositorios seguros, con acceso restringido y trazabilidad completa. El principio legal es contundente: lo que no se documenta, no existe. Una evidencia bien preservada puede ser la diferencia entre demostrar una intrusión maliciosa y quedar indefensos ante las autoridades o aseguradoras. Además, el análisis posterior de estos incidentes permite fortalecer los controles y prevenir futuros ataques.

9.6 Comunicación interna controlada

La comunicación durante una crisis debe ser precisa y estratégica. Un flujo de información desordenado o excesivo puede generar confusión, pánico o rumores, amplificando el daño del incidente. Por eso, se establecen lineamientos claros: solo se comunica lo mínimo necesario, los mensajes deben ser claros y libres de especulación, la vocería debe estar definida previamente y se debe proteger la identidad y privacidad de las personas involucradas.

El objetivo es doble: evitar la propagación de rumores que afecten la moral y la confianza, y preservar la continuidad operativa de la organización. Comunicar menos y mejor protege mucho más que tratar de explicar todo rápidamente, ya que permite controlar la narrativa y evitar interpretaciones erróneas o malintencionadas.

Finalmente, la correcta activación y ejecución del protocolo de respuesta inmediata transforma una amenaza activa en un evento controlado. La organización demuestra así que no improvisa ni actúa bajo miedo, sino que cuenta con liderazgo, método y respaldo institucional, protegiendo lo más valioso: la capacidad de decidir con claridad bajo presión y mantener la resiliencia ante amenazas sofisticadas.

10. Gestión de Crisis y Continuidad del Negocio

Cuando un incidente de deepfake en tiempo real escala a nivel de crisis, la organización entra en un escenario radicalmente distinto al de la operación normal. En este punto, la prioridad deja de ser exclusivamente técnica o preventiva y se transforma en una responsabilidad integral de liderazgo: preservar la continuidad del negocio, la reputación institucional y la autoridad del gobierno corporativo.

Una crisis de deepfake no solo amenaza activos financieros o sistemas tecnológicos; amenaza la confianza, tanto interna como externa. La credibilidad del liderazgo, la solidez de los procesos y la capacidad de tomar decisiones bajo presión quedan expuestas en tiempo real ante empleados, clientes, socios, reguladores y opinión pública.

Este capítulo establece cómo una organización madura gobierna la crisis, en lugar de reaccionar a ella. No se trata de improvisar soluciones, sino de activar una arquitectura de control diseñada previamente para operar en contextos de alta incertidumbre.

La diferencia entre una empresa que sobrevive a una crisis y una que queda marcada durante años no radica en evitar el incidente, sino en cómo dirige el caos, cómo comunica bajo presión y cómo protege su continuidad estratégica.

Principio rector

La crisis no revela debilidades; revela si el liderazgo estaba preparado.

Una organización preparada no entra en pánico, no fragmenta decisiones ni pierde el control narrativo. Activa estructuras claras, asume responsabilidad institucional y mantiene el rumbo estratégico aun en entornos hostiles.

10.1 Comité de crisis

Rol estratégico

El Comité de Crisis es el órgano máximo de decisión temporal durante un incidente crítico. Su función no es operativa, sino estratégica y directiva. Durante la crisis, este comité sustituye los flujos normales de gobierno, centralizando autoridad, criterio y comunicación.

El objetivo principal del comité es:

- Reducir la dispersión de decisiones.
- Evitar contradicciones internas y externas.
- Alinear respuesta técnica, legal, reputacional y de negocio bajo una sola narrativa y un solo mando.

En un escenario de deepfake en tiempo real, donde la información cambia minuto a minuto, la fragmentación decisional es uno de los mayores riesgos. El Comité de Crisis existe precisamente para neutralizar ese riesgo.

Integración mínima

El comité debe estar integrado, como mínimo, por las siguientes funciones clave:

- **CEO (líder del comité):**
  Responsable último de la toma de decisiones estratégicas y de representar la autoridad institucional. Su liderazgo es simbólico y operativo.
- **Seguridad Corporativa:**
  Aporta análisis de riesgo, evaluación de impacto, preservación de evidencia y coordinación con fuerzas externas cuando sea necesario.
- **Jurídico / Compliance:**
  Evalúa implicaciones legales, regulatorias y contractuales, valida mensajes públicos y protege a la organización de autoincriminación.
- **Comunicación Corporativa:**
  Diseña y ejecuta la estrategia narrativa, controla vocerías y monitorea impacto mediático y reputacional.
- **TI / Ciberseguridad:**
  Proporciona información técnica, contención del incidente y evaluación de vectores de ataque.
- **Área directamente impactada:**
  Aporta conocimiento operativo del proceso afectado y facilita la implementación de medidas correctivas.

En incidentes de alta gravedad, el comité **informa y coordina directamente con el Consejo de Administración**, manteniendo un canal claro y documentado de comunicación.

Principios de operación

El Comité de Crisis debe operar bajo reglas estrictas:

- **Autoridad clara y concentrada:**
  No hay decisiones paralelas ni “autonomías locales” durante la crisis.
- **Reuniones breves, frecuentes y documentadas:**
  La cadencia importa más que la duración. Todo debe quedar registrado.
- **Decisiones colegiadas, ejecución centralizada:**
  Se escucha a todos los expertos, pero se ejecuta bajo un solo mando.

Mensaje ejecutivo

En crisis, menos voces y más claridad.

10.2 Toma de decisiones bajo incertidumbre

Realidad operativa

Un incidente de deepfake en tiempo real se caracteriza por un entorno hostil a la toma de decisiones racionales:

- La información es incompleta o contradictoria.
- El tiempo es limitado y los daños escalan rápidamente.
- La presión interna y externa es extrema.

En este contexto, esperar a tener certeza absoluta **paraliza la organización** y amplifica el impacto del incidente.

Principios decisionales

La toma de decisiones en crisis debe regirse por principios distintos a los de la operación normal:

- **Priorizar protección sobre optimización:**
  Es preferible una decisión conservadora que proteja a la organización que una decisión óptima pero tardía.
- **Decidir con información suficiente, no perfecta:**
  La suficiencia se define por umbrales de riesgo, no por exhaustividad.
- **Revisar y ajustar decisiones conforme evoluciona el contexto:**
  Las decisiones no son definitivas; son iterativas y adaptativas.

Herramientas clave

Para facilitar decisiones efectivas bajo presión, el comité debe apoyarse en:

- **Escenarios mínimos y máximos:**
  Definir el mejor y el peor escenario posible para anticipar impactos.
- **Límites claros de pérdida aceptable:**
  Determinar previamente qué nivel de daño es tolerable y cuál no.
- **Registro formal de decisiones:**
  Documentar qué se decidió, por qué y con qué información disponible.

Mensaje clave

No decidir también es una decisión, y suele ser la peor.

10.3 Manejo reputacional

Riesgo central

En incidentes de deepfake, el daño reputacional suele superar al daño financiero. La percepción pública de descontrol, negligencia o falta de liderazgo puede erosionar la confianza de manera irreversible.

La reputación no se protege ocultando información, sino gestionando la narrativa con coherencia y responsabilidad.

Estrategia

Una estrategia reputacional efectiva debe:

- Reconocer hechos confirmados sin especular.
- Proteger la credibilidad del liderazgo institucional.
- Evitar contradicciones públicas entre áreas o voceros.

El silencio prolongado o la negación temprana suelen interpretarse como culpa o incompetencia.

Errores comunes

Entre los errores más frecuentes en crisis de deepfake destacan:

- Negar hechos sin información suficiente.
- Buscar culpables individuales en lugar de asumir responsabilidad institucional.
- Comunicar tarde, de forma reactiva o fragmentada.

Principio

La reputación se pierde en minutos y se reconstruye en años.

10.4 Comunicación externa y medios

Vocería

La comunicación externa durante una crisis debe ser estrictamente controlada:

- Vocero único autorizado.
- Mensajes previamente validados por el área Legal.
- Coordinación con el Consejo de Administración cuando aplique**.**

Cualquier mensaje no autorizado aumenta el riesgo reputacional y legal.

Contenido del mensaje

Todo mensaje externo debe responder, de forma clara y consistente, a cuatro preguntas:

- Qué pasó (solo hechos confirmados).

- Qué se está haciendo para contener y corregir.
- Qué controles existen para evitar recurrencia.
- Compromiso explícito con la transparencia y la responsabilidad.

Canales

Los canales deben seleccionarse estratégicamente:

- Comunicados oficiales.
- Notificación a autoridades regulatorias.
- Medios estratégicos previamente definidos.

Mensaje rector

Hablar con orden protege más que guardar silencio desorganizado.

10.5 Coordinación con autoridades

Cuándo coordinar

La coordinación con autoridades es obligatoria cuando el incidente implica:

- Impacto financiero relevante.
- Riesgo o exposición de datos personales.
- Posibles implicaciones penales o regulatorias.

Retrasar esta coordinación agrava el riesgo legal.

Autoridades típicas

Dependiendo del caso, se coordina con:

- Fiscalía correspondiente.
- Autoridades financieras.
- INAI, cuando existe impacto en datos personales.

Principios

La relación con autoridades debe basarse en:

- Cooperación sin autoincriminación.
- Preservación estricta de evidencia.
- Acompañamiento legal permanente.

Mensaje clave

No coordinar a tiempo convierte un incidente en un problema legal mayor.

10.6 Recuperación post-incidente

Objetivos

Una vez contenida la crisis, la organización debe enfocarse en:

- Restablecer la operación normal.
- Reparar la confianza interna y externa.
- Fortalecer controles y capacidades.

Acciones clave

La fase de recuperación incluye:

- Análisis post-mortem sin búsqueda de culpables.
- Ajuste de políticas, protocolos y controles.
- Apoyo psicológico y organizacional al personal afectado.
- Reporte formal y estructurado al Consejo de Administración.

Medida de éxito

El éxito no se mide por la ausencia futura de incidentes, sino por:

Que si vuelve a pasar, el impacto sea menor y la respuesta mejor.

Conclusión Ejecutiva

La gestión de crisis es la prueba definitiva del gobierno corporativo. Este capítulo demuestra que una organización preparada no improvisa liderazgo, no pierde control narrativo y no sacrifica continuidad por miedo.

Capítulo 11. Tecnología y Herramientas de Apoyo en la Gestión de Crisis por Deepfakes en Tiempo Real

Introducción: Importancia Estratégica de la Tecnología en la Gestión de Crisis por Deepfakes

En el contexto actual, donde los ataques mediante deepfakes en tiempo real representan una amenaza creciente y sofisticada para la integridad de las organizaciones, la tecnología se erige como un habilitador esencial en la gestión de crisis. Sin embargo, es crucial comprender que la tecnología, por sí sola, no constituye una garantía absoluta de

protección. La dependencia acrítica de herramientas tecnológicas puede incrementar el riesgo tanto como ignorarlas.

Este capítulo profundiza en las funciones, limitaciones y recomendaciones de los principales recursos tecnológicos disponibles, orientando a la alta dirección y al consejo de administración sobre cómo integrarlos de manera estratégica al gobierno corporativo, evitando falsas expectativas y preservando la capacidad de decisión bajo presión.

11.1 Sistemas de Detección de Voz Sintética

Funcionamiento

Los sistemas de detección de voz sintética emplean algoritmos avanzados de inteligencia artificial y análisis de señales acústicas para identificar patrones no humanos en el audio. Estos sistemas analizan características como la consistencia tonal, prosódica, frecuencia, ritmo y respiración, buscando artefactos generados por modelos de IA que suelen ser imperceptibles para el oído humano. El proceso puede ser realizado en tiempo real o en modo forense post-incidente, dependiendo de la infraestructura y necesidades operativas.

Capacidades Actuales

- ✓ Reconocimiento de artefactos acústicos propios de la síntesis por IA.
- ✓ Análisis de variaciones en tonalidad, ritmo y pausas que difieren de la voz humana natural.
- ✓ Detección de anomalías en la cadencia respiratoria y en la modulación de la voz.
- ✓ Integración con sistemas de monitoreo de llamadas en centros críticos y plataformas de comunicación institucional.
- ✓ Casos de Uso Válidos
- ✓ Monitoreo de llamadas en tiempo real en áreas sensibles (finanzas, operaciones críticas, atención a clientes VIP).
- ✓ Análisis forense de grabaciones ante sospecha de suplantación de identidad o fraude.
- ✓ Validación de instrucciones de alto impacto en procesos de autorización y toma de decisiones.

Riesgos y Recomendaciones

Falsos negativos: Los modelos de IA evolucionan rápidamente, lo que puede llevar a que algunos deepfakes avanzados no sean detectados, exponiendo a la organización a riesgos latentes.

Falsos positivos: Una detección errónea puede paralizar la operación y generar disputas internas, afectando la confianza en los sistemas.

Dependencia sin criterio humano: La automatización excesiva sin supervisión experta puede llevar a la relajación de controles críticos y a la pérdida de la capacidad de discernimiento ante situaciones complejas.

Recomendación: Integrar los sistemas de detección como herramientas de apoyo, nunca como sustitutos del juicio humano. Capacitar al personal en la interpretación de resultados, y establecer protocolos de doble validación en casos críticos.

11.2 Plataformas Seguras de Comunicación

- ✓ Características Indispensables
- ✓ Autenticación multifactor para garantizar la identidad de los participantes.
- ✓ Registro y trazabilidad completa de las interacciones, con logs auditables.
- ✓ Integridad de mensajes mediante cifrado de extremo a extremo y mecanismos de verificación.
- ✓ Control de acceso granular por rol y nivel de autorización.
- ✓ Capacidad de verificación fuera de canal para instrucciones sensibles.

Beneficios Estratégicos

- ✓ Reducción significativa de la superficie de ataque, limitando el acceso a canales institucionales controlados.
- ✓ Evidencia auditable para respaldar investigaciones y procesos legales.
- ✓ Disminución de disputas internas y externas al contar con registros verificables.
- ✓ Mitigación de riesgos asociados a suplantación y manipulación de mensajes.

Implementación y Relevancia en la Prevención de Ataques

La implementación de plataformas seguras debe ser un proyecto transversal, involucrando áreas de TI, seguridad, jurídico y operaciones. Es crucial realizar auditorías periódicas, simulaciones de ataque y capacitaciones continuas al personal directivo y operativo. Un canal inseguro convierte cualquier instrucción en una amenaza potencial, por lo que la revisión constante de la arquitectura y los protocolos de comunicación es indispensable.

11.3 Análisis Biométrico Conductual

Concepto

El análisis biométrico conductual se centra en el estudio de las acciones y patrones de comportamiento de las personas, en lugar de sus características físicas o vocales. Utiliza datos históricos y algoritmos de aprendizaje automático para identificar desviaciones respecto al patrón habitual de interacción.

Variables Analizadas

Ritmo y frecuencia de interacción con sistemas y plataformas.

Patrones de escritura, velocidad, estructura gramatical y sintaxis.

Secuencia de toma de decisiones en procesos críticos.

Reacciones ante estímulos inesperados o situaciones de presión.

Ventajas

Dificultad de clonación en tiempo real, lo que complica la suplantación por parte de deepfakes.

Complemento efectivo a controles físicos y tecnológicos, aumentando la robustez del sistema de defensa.

Capacidad de detección proactiva de intentos de manipulación o acceso fraudulento.

Limitaciones y Aplicaciones Prácticas

Requiere históricos amplios y fiables para establecer patrones significativos.

Es sensible a cambios contextuales, como estrés, fatiga o alteraciones en el entorno laboral.

Debe ser aplicado en combinación con otros controles para obtener resultados óptimos.

Aplicaciones prácticas: Validación de accesos privilegiados, monitoreo de actividades en puestos críticos, apoyo en investigaciones internas y fortalecimiento de procesos de autenticación.

11.4 Limitaciones Tecnológicas Actuales

Desafíos y Riesgos Estratégicos

No existe solución tecnológica capaz de detectar todos los deepfakes en tiempo real con precisión absoluta.

La velocidad de evolución de los modelos de ataque supera a la capacidad de adaptación de las defensas tecnológicas.

La sensación falsa de seguridad puede llevar a la relajación de controles humanos y a la delegación excesiva en sistemas automatizados.

La dependencia excesiva de proveedores externos pone en riesgo la soberanía tecnológica y la capacidad de respuesta ante incidentes.

Evolución de Amenazas y Necesidad de Controles Humanos

El panorama de amenazas cambia constantemente, obligando a las organizaciones a mantener una postura activa de revisión y actualización de sus sistemas. La tecnología debe ser vista como un aliado, nunca como un reemplazo del criterio y la supervisión humana. La gobernanza efectiva requiere la integración de controles manuales y tecnológicos, así como la formación continua de los equipos responsables.

11.5 Evaluación y Selección de Proveedores

Criterios para la Alta Dirección

- Transparencia metodológica: El proveedor debe explicar los métodos y algoritmos utilizados, evitando soluciones “caja negra”.
- Evidencias de pruebas reales y resultados auditados por terceros independientes.

- Claridad sobre tasas de error, limitaciones y escenarios de aplicación.
- Facilidad de integración con procesos humanos y sistemas existentes.
- Soporte legal y contractual robusto, incluyendo cláusulas de responsabilidad y auditoría.
- Red Flags y Recomendaciones
- Promesas absolutas de detección o protección.
- Ausencia de auditorías independientes o validaciones externas.
- Opacidad técnica, falta de documentación o de acceso a información relevante.
- Falta de soporte ante incidentes o de compromiso contractual con la mejora continua.

Recomendación: Involucrar al área jurídica y de compliance en la selección de proveedores, exigir pruebas piloto y establecer mecanismos de revisión periódica de desempeño y seguridad.

Conclusión Ejecutiva

La tecnología es un componente indispensable en la defensa contra deepfakes en tiempo real; sin embargo, su sola presencia no constituye una solución definitiva. La verdadera ventaja competitiva radica en la capacidad de gobernar los recursos tecnológicos, comprendiendo sus alcances, limitaciones y riesgos. La alta dirección y el consejo de administración deben priorizar la integración estratégica, la capacitación continua y la supervisión activa, asegurando que la toma de decisiones bajo presión se apoye en herramientas robustas, pero nunca dependa exclusivamente de ellas.

La organización que reconoce y gestiona los límites de la tecnología es la que mejor protege su capacidad de decidir y de responder eficazmente ante crisis, fortaleciendo así el gobierno corporativo y la resiliencia institucional.

Capítulo 12. Auditoría, Monitoreo y Mejora Continua: Gobernanza Estratégica ante Deepfakes en Tiempo Real

Manual Ejecutivo para Alta Dirección y Consejo de Administración

Introducción

En el entorno corporativo contemporáneo, caracterizado por la acelerada evolución de las amenazas digitales, la auditoría, el monitoreo y la mejora continua constituyen el núcleo de una defensa efectiva ante el fenómeno de los deepfakes en tiempo real. Para la Alta Dirección y el Consejo, estos procesos no son simples formalidades regulatorias, sino instrumentos estratégicos que garantizan la capacidad de respuesta, la preservación de la reputación y el cumplimiento fiduciario. La resiliencia institucional frente a amenazas emergentes depende de una gobernanza activa, capaz de detectar, corregir y anticipar riesgos, transformando cada incidente en una oportunidad de fortalecimiento organizacional.

12.1 Indicadores Clave (KPIs): Medición Estratégica y Toma de Decisiones Informada

Los KPIs son la herramienta fundamental para traducir la percepción de seguridad en métricas objetivas, permitiendo a la Alta Dirección y al Consejo visualizar el verdadero nivel de preparación ante deepfakes. Una gestión basada en indicadores evita la complacencia y orienta la toma de decisiones bajo criterios de evidencia.

Definición y Categorías:

Indicadores de Prevención: Miden la capacidad proactiva de la organización para evitar incidentes. Ejemplo: porcentaje de personal crítico capacitado en reconocimiento de deepfakes, número de procesos con doble validación activa, cobertura de verificación multicanal.

Indicadores de Detección: Evalúan la rapidez y eficacia en la identificación de anomalías. Ejemplo: tiempo promedio de detección, número de sospechas reportadas (incluyendo falsos positivos), porcentaje de reportes oportunos vs. tardíos.

Indicadores de Respuesta: Reflejan la capacidad de activar protocolos y aislar decisiones ante una amenaza. Ejemplo: tiempo de activación del protocolo, tiempo de aislamiento de decisiones, cumplimiento de escalamiento jerárquico.

Indicadores Culturales: Revelan el nivel de confianza interno y la disposición a reportar incidentes. Ejemplo: resultados de encuestas post-simulacro, incidentes no reportados detectados en auditoría.

Ejemplos Prácticos:

Un aumento en el número de reportes puede indicar una cultura madura de alerta temprana, no necesariamente debilidad operativa.

La reducción del tiempo de reacción tras simulacros evidencia la eficacia de la capacitación y la claridad de roles.

Interpretación y Impacto:

KPIs deben ser revisados periódicamente en sesiones de Consejo, integrando la información en la estrategia corporativa.

Las desviaciones significativas requieren análisis inmediato y posible ajuste de protocolos.

12.2 Auditorías Internas: Validación Real y Seguimiento Estratégico

La auditoría interna es el mecanismo que transforma los controles de papel en prácticas efectivas, validando la ejecución bajo condiciones reales y asegurando la preparación ante deepfakes. Su rol estratégico es doble: prevenir errores futuros y fortalecer la gobernanza.

Metodología y Alcance Ampliado:

Revisión integral de cumplimiento de protocolos anti-deepfake, desde la capacitación hasta la respuesta operativa.

Evaluación de segregación de funciones para evitar concentración de decisiones y minimizar riesgos de manipulación.

Verificación de trazabilidad digital y documentación de evidencias, asegurando la capacidad de reconstruir incidentes.

Revisión de contratos con terceros, garantizando que proveedores clave cumplan con estándares de seguridad y auditoría.

Mejores Prácticas:

Enfoque sistémico, orientado a procesos y no a personas, promoviendo la mejora continua en lugar de la sanción.

Reporte directo al Consejo, favoreciendo la transparencia y la acción ejecutiva oportuna.

Uso de auditorías sorpresa para evaluar la resiliencia real bajo presión.

Seguimiento y Evaluación:

Elaboración de informes ejecutivos que incluyan hallazgos, recomendaciones, y planes de acción concretos con responsables y fechas de cumplimiento.

Monitoreo de avances y cierre de brechas detectadas en auditorías previas.

12.3 Pruebas y Simulacros: Preparación Dinámica y Cultura de Resiliencia

Las pruebas y simulacros permiten comprobar la capacidad de reacción institucional ante la materialización de un ataque con deepfake, alineando la percepción con la realidad operativa. Son el laboratorio donde se ensaya la toma de decisiones bajo presión.

Tipos y Diseño:

Simulacros de escritorio (table-top): análisis teórico de escenarios, ideal para la Alta Dirección.

Ejercicios sorpresa controlados: activación inesperada de protocolos, evaluando la capacidad de respuesta inmediata.

Simulaciones específicas para la alta dirección: evaluación del criterio decisorio ante situaciones ambiguas.

Pruebas con terceros clave: integración de proveedores y aliados estratégicos en la respuesta conjunta.

Ejecución y Evaluación:

Definición clara de roles y responsabilidades previos a cada simulacro.

Medición de tiempos de reacción, calidad de las decisiones tomadas y coordinación interáreas.

Documentación de aprendizajes y áreas de mejora en reportes post-simulacro.

Integración con la Cultura Organizacional:

Los simulacros deben ser frecuentes y variados para evitar la rutina y promover la adaptabilidad.

Su ejecución debe comunicarse como un ejercicio de fortalecimiento, no de evaluación punitiva.

12.4 Actualización del Manual: Dinamismo, Gobernanza y Gestión del Cambio

El manual de defensa ante deepfakes debe ser un documento vivo, reflejo de la evolución tecnológica, regulatoria y de amenazas emergentes. La actualización constante es clave para evitar la obsolescencia y la falsa sensación de seguridad.

Proceso de Revisión:

Revisión mínima anual, alineada con el calendario de auditoría interna y los reportes de simulacros.

Actualización inmediata tras incidentes relevantes, incorporando lecciones aprendidas y ajustes operativos.

Ajustes ante cambios regulatorios, tecnológicos o de contexto de negocio.

Gobernanza y Gestión del Cambio:

Liderazgo de Seguridad Corporativa, con validación de las áreas Legal y Auditoría.

Aprobación final por la Alta Dirección y el Consejo, asegurando alineación estratégica.

Comunicación institucional clara sobre los cambios implementados y su impacto en la operación.

Riesgos Asociados:

Un manual desactualizado es fuente de falsa confianza y vulnerabilidad, especialmente ante amenazas dinámicas como los deepfakes.

12.5 Lecciones Aprendidas: Aprendizaje Rigurosos y Fortalecimiento Institucional

La captura y análisis de lecciones aprendidas es el motor de la mejora continua y el blindaje institucional ante la repetición de errores. Este proceso debe ser sistemático, libre de culpas y orientado al fortalecimiento operativo.

Proceso de Captura y Análisis:

Recopilación estructurada de incidentes, simulacros y auditorías, identificando qué funcionó, qué falló y qué señales se ignoraron.

Análisis objetivo de decisiones retrasadas y factores que influyeron en la respuesta.

Integración de hallazgos en protocolos, reentrenamiento focalizado y comunicación institucional.

Ejemplos de Aplicación:

Tras un ataque deepfake no detectado a tiempo, ajuste inmediato de los criterios de validación y capacitación dirigida a los equipos involucrados.

Incorporación de nuevos indicadores y simulacros en función de patrones emergentes identificados en incidentes recientes.

Comunicación Institucional:

Difusión de aprendizajes relevantes para toda la organización, fortaleciendo la cultura de transparencia y mejora continua.

Conclusión Ejecutiva

La auditoría, el monitoreo y la mejora continua transforman la gestión de riesgos ante deepfakes en tiempo real en un sistema activo de resiliencia corporativa. Para la Alta Dirección y el Consejo, la verdadera protección no reside en la perfección de los controles, sino en la capacidad organizada de aprender, adaptarse y anticipar amenazas, superando la velocidad del atacante. La resiliencia institucional se construye con disciplina, rigor y liderazgo estratégico, y se prueba en cada auditoría, simulacro y actualización del manual. Así, la organización se posiciona no solo para resistir, sino para prosperar en escenarios de crisis digital.

13. Capacitación y Cultura Organizacional

Pilares Estratégicos de Resiliencia ante Deepfakes en Tiempo Real

La gestión de riesgos ante deepfakes en tiempo real trasciende la implementación tecnológica: reside en la fortaleza de la cultura organizacional y la solidez de los programas de capacitación. El éxito de los controles depende, en última instancia, de la capacidad de la organización para anticipar, detectar y resistir intentos de manipulación sofisticada. Por ello, la cultura y la capacitación no son actividades accesorias, sino activos estratégicos que habilitan la resiliencia institucional. Este capítulo describe en detalle cómo la Alta Dirección y el Consejo pueden construir y sostener una cultura organizacional que, lejos de ser complaciente, se mantiene atenta, cuestionadora y adaptable frente a amenazas en constante evolución.

13.1 Programas para Directivos

Diagnóstico y Amenazas Específicas: Los directivos y consejeros son los objetivos predilectos de los ataques con deepfakes en tiempo real debido al valor de sus decisiones, su exposición mediática y el acceso privilegiado a información crítica. Los

intentos de suplantación, manipulación de decisiones estratégicas y explotación de la cadena de mando son vectores recurrentes. No basta con la conciencia general; se requiere una capacitación diferenciada y adaptada al perfil de riesgos y responsabilidades de la Alta Dirección.

Objetivos Estratégicos del Programa:

Asumir el deepfake como un riesgo de negocio, no solo como una amenaza tecnológica.

Desarrollar la habilidad de liderar procesos de verificación, estableciendo la validación como una práctica de liderazgo y no como un acto de desconfianza.

Tomar decisiones informadas bajo presión, incertidumbre y en contextos de potencial manipulación emocional.

Reconocer patrones conductuales y señales sutiles de ingeniería social dirigida a su rol.

Comprender las implicaciones legales, reputacionales y fiduciarias de sus acciones ante incidentes de deepfake.

Contenidos Específicos:

Evolución de los ataques deepfake en tiempo real: Análisis de casos emblemáticos a nivel internacional, impacto en consejos y alta dirección, lecciones aprendidas.

Psicología del engaño para líderes: Estrategias de manipulación emocional, explotación de la urgencia y técnicas de persuasión utilizadas por atacantes.

Suplantación jerárquica y abuso de autoridad: Cómo identificar intentos de falsificación de identidad en canales formales e informales.

Responsabilidad fiduciaria: Obligaciones legales, impacto sobre la cadena de custodia de decisiones y consecuencias de la omisión de controles.

Gestión de incidentes activos: Rol del directivo en la contención, comunicación y reporte oportuno de eventos sospechosos o confirmados.

Formatos y Metodología:

Sesiones cerradas y confidenciales para el Consejo y C-Suite, enfocadas en análisis de escenarios y discusiones estratégicas.

Talleres interactivos con simulaciones personalizadas y ejercicios de toma de decisiones bajo presión.

Briefings ejecutivos periódicos sobre tendencias emergentes, actualizaciones regulatorias y nuevos vectores de ataque.

Materiales de referencia ejecutiva y protocolos de actuación rápida, diseñados para consulta inmediata.

Desarrollo de Liderazgo Verificativo: El liderazgo efectivo ante deepfakes exige un cambio cultural: verificar la autenticidad de la información y la identidad de los interlocutores es un deber estratégico, no una señal de debilidad o desconfianza. La Alta Dirección debe modelar y promover esta conducta en toda la organización.

13.2 Entrenamiento Operativo

Alcance y Segmentación: El personal operativo representa la primera línea de defensa ante deepfakes, ya que interactúa directamente con procesos críticos y clientes internos y externos. El entrenamiento debe ser específico para áreas de mayor exposición y riesgo: finanzas, tesorería, pagos, compras, recursos humanos, asistentes ejecutivos, seguridad, TI y centros de monitoreo.

Contenidos Prácticos y Relevantes:

Definición operativa de deepfake: Qué son, cómo operan, diferencias con fraudes tradicionales y ejemplos concretos por área.

Canales de ataque: Identificación de intentos en llamadas telefónicas, videollamadas, mensajes de voz, correos y plataformas colaborativas.

Indicadores técnicos y comportamentales: Detección de irregularidades en la voz, imagen, lenguaje corporal, patrones de urgencia o cambios en rutinas habituales.

Protocolos de actuación: Pasos inmediatos ante sospecha, canales de reporte, escalamiento seguro y criterios para detener procesos potencialmente comprometidos.

Errores críticos por evitar: Acciones prohibidas, incluso si la instrucción parece legítima, y casos históricos de fallos por exceso de confianza.

Principios Pedagógicos:

Lenguaje claro, directo y adaptado al nivel de conocimiento del público; evitar tecnicismos innecesarios.

Uso intensivo de casos reales, testimonios internos y simulaciones contextualizadas al entorno de la organización.

Sesiones recurrentes y formatos breves para mantener la vigencia del aprendizaje y adaptarse a patrones emergentes.

Evaluaciones prácticas mediante ejercicios de toma de decisiones, con retroalimentación inmediata y orientada a la mejora.

Ejemplos de Aplicación:

Simulación de una instrucción urgente de pago proveniente de un directivo falso, con análisis de la respuesta y decisión tomada.

Detección de una videollamada manipulada y activación del protocolo de escalamiento seguro.

Reporte y bloqueo de un intento de suplantación en la contratación de personal.

Evaluación y Certificación: El entrenamiento debe incluir indicadores de desempeño, seguimiento individual y certificación periódica como requisito de permanencia en procesos críticos. El objetivo no es la obediencia ciega, sino la formación de criterios informados y responsables.

13.3 Simulaciones de Fraude

Justificación Estratégica: Las simulaciones de fraude con deepfakes son el método más efectivo para validar la preparación real de la organización, identificar vulnerabilidades humanas y fortalecer la cultura de reporte y verificación. Su diseño debe ser riguroso, realista y orientado a la mejora continua, no a la sanción.

Tipos de Simulación:

Llamadas falsas de supuestos directivos, con órdenes urgentes o inusuales.

Videollamadas manipuladas en tiempo real, evaluando la capacidad de identificar inconsistencias visuales o de comportamiento.

Mensajes apremiantes solicitando pagos, transferencias o liberación de información sensible.

Escenarios de crisis reputacional, donde se evalúa la reacción ante comunicaciones oficiales adulteradas.

Metodología y Ejecución:

Simulacros sin aviso previo, diseñados para medir la reacción espontánea y la adherencia a los protocolos establecidos.

Evaluación estructurada de respuestas, incluyendo tiempos de reacción, canales utilizados y calidad del reporte.

Retroalimentación inmediata y personalizada a los equipos involucrados, con foco en el aprendizaje y no en la penalización.

Registro formal de hallazgos, análisis de patrones de error y ajustes inmediatos a procesos y controles.

Impacto en la Cultura Organizacional: Las simulaciones normalizan la práctica de la verificación, promueven la transparencia y desmitifican el error como fuente de aprendizaje. Refuerzan la importancia de reportar dudas y de actuar con rigor, incluso bajo presión. Simular no es desconfiar: es construir una cultura resiliente y preparada.

13.4 Integración con Protocolos y Mejora Continua

El valor estratégico de la capacitación y las simulaciones reside en su capacidad de retroalimentar, en tiempo real, la actualización de protocolos, la definición de nuevos indicadores y el rediseño de procesos críticos. Toda lección aprendida debe traducirse en

acciones concretas: ajuste de criterios de validación, rediseño de flujos de autorización, actualización de manuales, y reentrenamiento focalizado. Es fundamental comunicar los aprendizajes de manera transversal, incentivar el reporte sin temor a represalias y reconocer públicamente la detección temprana como un logro organizacional.

Las áreas responsables de auditoría, riesgos y recursos humanos deben coordinarse estrechamente para garantizar que los hallazgos se integren a los planes de capacitación, que los incidentes se documenten de forma estructurada y que la comunicación institucional refuerce la cultura de mejora continua.

13.5 Cultura de Reporte sin Represalias

Una cultura que castiga el error incentiva el silencio, y el silencio es el mayor aliado del fraude.

Principios culturales

- Reportar sospechas es obligatorio y protegido.
- No se penaliza la duda razonable.
- Se reconoce públicamente la detección temprana.

Mecanismos formales

- Canales claros de reporte.
- Escalamiento sin fricción.
- Protección contra represalias.
- Confidencialidad garantizada.

Mensaje inequívoco desde la Alta Dirección:

Prefiero cien alertas falsas que un solo fraude no reportado.

13.6 Evaluación de Madurez Organizacional

La cultura se mide. La capacitación se audita. La madurez se evalúa.

Dimensiones de evaluación

- Conocimiento del riesgo.
- Adherencia a protocolos.
- Velocidad de reporte.
- Comportamiento bajo presión.

- Nivel de confianza psicológica.

Niveles de madurez

1. Reactivo: Desconocimiento y negación.
2. Consciente: Capacitación básica.
3. Estructurado: Protocolos aplicados.
4. Integrado: Cultura de verificación.
5. Resiliente: Anticipación y aprendizaje continuo.

Uso estratégico

- Reporte a Consejo.
- Priorización de inversiones.
- Ajuste de controles.
- Benchmark interno y externo.

Conclusión Ejecutiva

La resiliencia organizacional ante deepfakes no se sostiene sobre la perfección tecnológica, sino sobre la madurez de la cultura y la disciplina en la capacitación. Una Alta Dirección que lidera con el ejemplo, equipos operativos entrenados para detectar y reportar, y simulaciones que desafían continuamente el statu quo, convierten el aprendizaje en un ciclo virtuoso de adaptación y anticipación. Así, la organización no solo resiste, sino prospera frente a escenarios de crisis digital, consolidando la cultura y la capacitación como sus activos estratégicos más valiosos.

14. Integración con Otros Sistemas

El riesgo de deepfakes en tiempo real no puede ni debe gestionarse como un silo. Cuando se aborda de forma aislada, genera duplicidades, vacíos de control y falsas sensaciones de seguridad. Cuando se integra correctamente, se convierte en un multiplicador de resiliencia organizacional.

Este capítulo establece cómo el Manual Corporativo Anti-deepfake se articula, sincroniza y refuerza con los principales sistemas de gestión existentes en la organización, asegurando coherencia estratégica, eficiencia operativa y trazabilidad para Alta Dirección y Consejo.

El principio rector es claro:

Un riesgo que no se integra, tarde o temprano se materializa.

14.1 Integración con Seguridad Corporativa

La Seguridad Corporativa es el eje natural de coordinación del riesgo deepfake, ya que combina visión estratégica, gestión de crisis y control operativo.

Rol de integración

- Incorporar el riesgo deepfake en el Modelo Integral de Seguridad.
- Alinear protocolos anti-deepfake con planes de protección ejecutiva, centros de monitoreo y gestión de incidentes.
- Garantizar que la detección de deepfakes tenga capacidad real de detención de decisiones.

Puntos clave de articulación

- Protocolos de crisis: el deepfake se trata como incidente de alta criticidad.
- Centros de Control: inclusión de alertas por suplantación de identidad.
- Seguridad ejecutiva: protección de imagen, voz y hábitos de comunicación del C-Suite.

Mensaje estratégico:

La seguridad ya no solo protege personas e instalaciones; protege decisiones.

14.2 Integración con SHE / Prevención

Aunque tradicionalmente enfocado en seguridad física y salud ocupacional, el sistema SHE comparte un principio esencial con el riesgo deepfake: la prevención basada en conducta y cultura.

Sinergias clave

- Cultura preventiva: identificar actos inseguros digitales como equivalentes a actos inseguros físicos.
- Reporte temprano: fomentar la notificación de anomalías sin temor.
- Gestión del estrés: preparación emocional para eventos críticos.

Analogía operativa

Así como nadie debe operar maquinaria sin validación, nadie debe ejecutar decisiones críticas sin verificación.

Mensaje cultural:

La prevención no distingue entre riesgos físicos y digitales; distingue entre preparación e improvisación.

14.3 Integración con Gestión de Riesgos

El deepfake en tiempo real debe figurar explícitamente dentro del Enterprise Risk Management (ERM) como riesgo emergente de alto impacto.

Integración estructural

- Clasificación como riesgo estratégico-operativo.
- Evaluación de impacto financiero, legal, reputacional y humano.
- Definición de apetito y tolerancia al riesgo.

Uso en toma de decisiones

- Priorización de inversiones.
- Asignación de recursos.
- Reportes periódicos a Consejo.

Principio clave:

Lo que no se mide como riesgo, se asume como sorpresa.

14.4 Integración con Compliance y Ética

El deepfake no solo es un riesgo operativo, es un riesgo ético y de cumplimiento.

Alineación normativa

- Códigos de conducta: obligación de verificación.
- Políticas anticorrupción: prevención de pagos indebidos inducidos por suplantación.
- Canales de denuncia: inclusión explícita de fraude por IA.

Dimensión ética

- Protección de la dignidad y reputación de directivos y colaboradores.
- Uso responsable de tecnologías de IA.
- Transparencia en la gestión de incidentes.

Mensaje institucional:

La ética no se suspende en una videollamada urgente.

14.5 Integración con Continuidad del Negocio

Un incidente de deepfake mal gestionado puede interrumpir operaciones críticas, generar decisiones irreversibles o desencadenar crisis reputacionales.

Articulación con BCP

- Inclusión de deepfakes como escenario disruptivo.
- Definición de decisiones críticas que deben pausarse ante sospecha.
- Roles claros durante continuidad y crisis.

Objetivo final

Garantizar que, incluso bajo ataque, la organización:

- Mantenga control decisorio.
- Proteja activos críticos.
- Preserve confianza de stakeholders.

Mensaje de resiliencia:

La continuidad del negocio depende de la continuidad del juicio.

Conclusión Ejecutiva

La integración transforma el Manual Anti-deepfake de un documento especializado en un sistema vivo de gobierno corporativo.

Cuando el riesgo deepfake se integra:

- La seguridad se vuelve estratégica.
- La prevención se vuelve cultural.
- El riesgo se vuelve visible.
- El cumplimiento se vuelve defensivo.
- La continuidad se vuelve resiliente.

Una organización integrada no elimina el riesgo, pero elimina la improvisación.

15. Anexos

Los anexos de este manual no son material secundario. Son herramientas ejecutables, diseñadas para operar bajo presión, reducir ambigüedad y convertir la estrategia en acción inmediata.

Este capítulo consolida los instrumentos que permiten que el Manual Corporativo Anti-deepfake funcione en la realidad, no solo en auditorías o presentaciones. Cada anexo está pensado para ser utilizado por directivos, mandos medios y personal crítico en tiempo real.

Principio rector:

En una crisis, nadie lee teorías; se ejecutan guías claras.

15.1 Glosario

El lenguaje común elimina errores críticos. Este glosario establece definiciones oficiales de uso corporativo.

Términos clave

- **Deepfake**: Contenido audiovisual generado o manipulado mediante inteligencia artificial para simular la identidad de una persona real.
- **Deepfake en tiempo real**: Suplantación activa de voz o imagen durante una interacción en vivo (llamada o videoconferencia).
- **Verificación multicanal**: Confirmación de identidad usando al menos dos canales independientes.
- **Canal seguro**: Medio previamente aprobado, autenticado y registrado por la organización.
- **Aislamiento de decisión**: Suspensión temporal de una acción crítica hasta completar verificación.
- **Incidente**: Evento confirmado de intento o ejecución de fraude.
- **Casi-incidente**: Evento detectado y contenido antes de causar impacto.
- **Escalamiento**: Comunicación formal a niveles superiores según protocolo.

Mensaje ejecutivo:

Definir términos es definir comportamientos.

15.2 Checklist Anti-deepfake

Este checklist debe utilizarse antes de ejecutar cualquier decisión crítica, especialmente bajo presión o urgencia.

Checklist esencial

- ☐ ¿La solicitud implica dinero, contratos, datos sensibles o reputación?
- ☐ ¿La instrucción se basa solo en voz o imagen?
- ☐ ¿Existe urgencia artificial o presión emocional?
- ☐ ¿El tono o lenguaje es inusual para la persona?
- ☐ ¿Se solicitó evitar validaciones normales?
- ☐ ¿Se verificó por un segundo canal seguro?
- ☐ ¿Se documentó la verificación?

Regla de oro:

Un solo "no verificado" detiene la decisión.

15.3 Flujogramas de Decisión

Los flujogramas eliminan improvisación y dependencia del criterio individual bajo estrés.

Flujo general de decisión

1. Recepción de solicitud.
2. Evaluación de criticidad.
3. Activación de checklist.
4. Verificación multicanal.
5. Escalamiento si existe duda.
6. Ejecución o bloqueo documentado.

Principios operativos

- Nadie es sancionado por detener una decisión.
- Toda duda razonable se escala.
- La velocidad nunca sustituye la verificación.

Mensaje operativo:

El flujo protege al colaborador tanto como a la empresa.

15.4 Formatos de Reporte

El reporte oportuno transforma un evento en aprendizaje organizacional.

Reporte de sospecha (mínimos)

- Fecha y hora.
- Canal utilizado.
- Identidad suplantada.
- Área impactada.
- Indicadores observados.
- Acción tomada.

Reporte post-incidente

- Descripción cronológica.
- Decisiones ejecutadas o detenidas.
- Impacto real o potencial.
- Controles que funcionaron o fallaron.
- Recomendaciones.

Principio legal:

Lo que no se documenta, no existe jurídicamente.

15.5 Roles y Contactos Críticos

La claridad de contactos reduce tiempos críticos.

Roles esenciales

- Líder del protocolo anti-deepfake
- Seguridad Corporativa
- TI / Ciberseguridad
- Legal / Compliance
- Comunicación Corporativa
- Alta Dirección designada

Reglas de contacto

- Información actualizada trimestralmente.
- Canales seguros predefinidos.

- Disponibilidad 24/7 para roles críticos.

Mensaje de cierre:

Un protocolo sin responsables es solo un documento.

Conclusión Ejecutiva

Estos anexos convierten la estrategia en acción inmediata.

Cuando se usan correctamente:

- Se reduce la carga cognitiva.
- Se protege al personal.
- Se acelera la respuesta.
- Se fortalece la defensa legal.

La diferencia entre una crisis contenida y un desastre corporativo suele estar en un anexo bien diseñado.

Bibliografias:

Chesney, R., & Citron, D. K. (2019). *Deepfakes and the new disinformation war: The coming age of post-truth geopolitics*. Foreign Affairs, 98(1), 147–155.

Citron, D. K., & Chesney, R. (2024). *Deepfakes and the law*. California Law Review, 112(1), 1–58. https://doi.org/10.2139/ssrn.3213954

Europol Innovation Lab. (2022). *Facing reality? Law enforcement and the challenge of deepfakes*. Europol.

Federal Bureau of Investigation. (2023). *Public service announcement: Deepfake-enabled fraud schemes*. FBI.

Gartner. (2023). *Emerging risks: AI-generated impersonation and executive fraud*. Gartner Research.

International Organization for Standardization. (2018). *ISO 31000: Risk management — Guidelines*. ISO.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). *Deepfakes: Trick or treat?* Business Horizons, 63(2), 135–146. https://doi.org/10.1016/j.bushor.2019.11.006

McKinsey & Company. (2024). *The risk implications of generative AI for organizations*. McKinsey Global Institute.

MIT Technology Review. (2023). *The rise of real-time deepfake fraud*. MIT Technology Review.

National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)*. NIST.

Organisation for Economic Co-operation and Development. (2022). *Global risk report: Emerging digital threats*. OECD.

PwC. (2023). *Global economic crime and fraud survey*. PwC.

Sullivan, B., & Hagen, J. (2023). *Voice cloning and biometric spoofing threats*. SANS Institute.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.

United Nations Office on Drugs and Crime. (2023). *Cybercrime and fraud enabled by artificial intelligence*. UNODC.

Verizon. (2024). *Data breach investigations report*. Verizon Enterprise.

World Economic Forum. (2024). *Global risks report*. World Economic Forum.